Trajectories of Success: A Longitudinal Analysis of Dynamic Factors Influencing Educational Achievement in Irish Children

Mohid Yasin

22201788

The thesis is submitted to University College Dublin
in part fulfilment of the requirements for the degree of

M.Sc. Statistics

School of Mathematics and Statistics

University College Dublin

Supervisor: Assistant Prof. Isabella Gollini

August 2024

Abstract

This longitudinal study investigates the key factors influencing cognitive development and educational achievement in Irish children using data from the Growing Up in Ireland study. Employing an innovative methodological approach that combines growth modelling techniques with advanced machine learning methods, the research captures both static characteristics and developmental trajectories of various factors affecting child outcomes. The study reveals that the trajectory of child behavioural issues over time is the strongest predictor of educational outcomes, surpassing initial behavioural status. Socioeconomic factors, particularly income and social class, demonstrate a persistent impact on educational achievement. Maternal education level emerges as a significant predictor, emphasising the intergenerational aspect of educational attainment. Notably, the interaction between changing behavioural patterns and initial socioeconomic status proves more predictive of educational outcomes than either factor alone. The research's novelty lies in its dynamic perspective, highlighting not just what factors are important, but how their influence evolves over time and in combination with other factors. This approach provides a more nuanced understanding of the complex interplay between various aspects of a child's life and environment in shaping educational outcomes. These findings have significant implications for educational policy and practice in Ireland, underscoring the need for early intervention programs targeting child behavioural issues and policies addressing socioeconomic disparities in education. The study also emphasizes the importance of supporting parental education and promoting sustained engagement in educational activities throughout childhood. While acknowledging limitations such as reliance on a single-point outcome measure and potential selection bias due to participant attrition, this research provides a robust evidence base for developing more effective, targeted interventions to support child development and educational achievement in Ireland.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Thesis Motivation

Children's ongoing and future development occurs in a wide range of topics, which include large domains such as their physical health and development, their social/emotional/behavioural wellbeing and their educational achievement, intellectual capacity and cognitive development.

Given these broad range of child outcomes, the thesis statement was in part motivated by identifying the topic that is cited by literature to be the most related to a child's continued development into the future.

While all aspects of child development are intertwined and are important, educational achievement stands out as particularly crucial for several reasons. Firstly, the long-term impact of educational achievement, particularly in childhood, is significant due to the cumulative nature of learning (Duncan et al., 2007). It forms the foundation for principles that lead to future learning. Strong early academic skills predict better academic performance throughout schooling and into higher education. Educational achievement in childhood has been consistently linked to better outcomes in the broader development of children into adulthood, including higher income, better health, and overall life satisfaction (Heckman, 2006).

Secondly, the transferability of cognitive and educational knowledge gained in early childhood is paramount. These skills are among the most transferable to various life domains, influencing problem-solving abilities, decision-making, and adaptability in changing environments(Ceci, 1991).

Lastly, the interrelation of educational achievement and cognitive development with other domains of child development is crucial. They can positively influence social-emotional well-being, behavioural outcomes, and even physical health through increased health literacy(Hair et al., 2015).

While acknowledging the importance of holistic child development, this thesis focuses on educational achievement and cognitive development due to their pervasive influence on a child's trajectory and the robust body of literature supporting their long-term significance.

1.2 Objectives

The primary objectives of this thesis are multifaceted. We aim to uncover significant associations by identifying and analysing the key factors that have the most substantial and statistically significant relationships with educational achievement and cognitive development in Irish children. This involves examining a wide range of variables across different domains of child development and their environment.

Another crucial objective is to identify timeline importance. We seek to determine the critical periods or age ranges where certain factors have the most pronounced impact on educational outcomes. This longitudinal perspective will also help in understanding how the trajectories of various factors may change or persist over time.

Contextualising our findings within the broader context of Irish society, education system, and policy landscape is also a key goal. This analysis may offer insights that could contribute to the broader global discourse on educational achievement and cognitive development, allowing for valuable cross-cultural comparisons and potentially identifying universal and culture-specific factors in educational achievement.

Furthermore, we aim to provide a baseline for future research by establishing a comprehensive foundation for future studies. By identifying those variables that demonstrate the most significant associations with educational achievement, we will help guide more targeted research efforts into those identified topics and potentially inform evidence-based interventions.

A secondary objective within this thesis is to conduct dropout analysis. We will examine the patterns and factors associated with participant attrition throughout the longitudinal study. This analysis aims to identify variables that explain differences between participants who completed all waves of the study and those who dropped out at various stages. By understanding these patterns, we can assess the potential for systematic dropout, a common challenge in longitudinal studies. This investigation will help evaluate the representativeness of the final sample, inform the interpretation of results, and provide insights for future longitudinal studies to minimize attrition and mitigate its effects on data quality and generalisability of findings.

1.3 Implications

The findings of this thesis are expected to have several important implications. In terms of educational policy, by identifying the most impactful factors related to educational achievement, this research can inform evidence-based policy making in Ireland. It may highlight areas where targeted interventions or resource allocation could yield the greatest benefits for children's educational outcomes.

Understanding the timeline importance of various factors can guide the development of early intervention strategies. This knowledge can help educators and policymakers focus on critical periods where specific interventions might be most effective.

By examining factors across various domains, including socioeconomic variables, this research may shed light on socioeconomic considerations, particularly issues of educational equity and social mobility in Ireland. This could inform broader social policies aimed at reducing educational disparities.

Given the interrelated nature of child development domains, this research may encourage more integrated approaches to supporting children's growth, potentially fostering cross-sector collaboration between education, health, and social services sectors.

Lastly, by focusing on educational achievement, which is linked to future economic outcomes, this research could have long-term economic implications for workforce development and economic planning in Ireland.

2 Data

2.1 Data Source: Growing Up in Ireland (GUI)

Growing Up in Ireland is Ireland's National Longitudinal Study of Children, commissioned in 2006 and funded by the Department of Health and Children through the Office of the Minister for Children and Youth Affairs, in association with the Department of Social Protection and the Central Statistics Office.

Key aspects of the GUI study include:

1. **Purpose:** To improve understanding of children's lives, promote their development, and give them a voice in matters affecting them.
2. **Design:** Longitudinal study with multiple cohorts, providing snapshots of children's lives at different points in time.
3. **Sample:** This thesis focuses on the Infant Cohort, which initially included 11,100 families.
4. **Waves:** While the study is ongoing, this thesis analyses data across separate timepoints known as "waves", which we will refer to from here onwards. Our data spanned 5 waves, from when the children were 9 months old to 9 years old.
5. **Areas of Focus:** Physical Health, Socio-Emotional Well-being, and Educational Achievement and Cognitive Development.

2.2 Data Characteristics and Relevance to Thesis

The Growing Up in Ireland (GUI) dataset provides an exceptional foundation for this thesis, offering both depth and breadth in its comprehensive approach to studying child development. Through extensive in-person interviews lasting an average of two hours and covering up to 400 questions per wave, the study captures a rich, multifaceted picture of children's and families' lives. The dataset's representativeness of the Irish population, coupled with its longitudinal design, enables the analysis of developmental trajectories and the identification of critical periods in children's educational and cognitive growth. This scale and depth create a unique opportunity to explore the complex interplay of factors affecting child development, aligning perfectly with this thesis's objectives. Ultimately, the GUI dataset offers a robust evidence base for both policy-making and future research, making it an ideal resource for uncovering significant associations and identifying the most impactful factors relating to educational achievement and cognitive development.

2.3 Approach to Data Visualisation and Additional Resources

Given the extensive use of the Growing Up in Ireland dataset, numerous comprehensive visualisations and analyses already exist. This thesis does not aim to replicate these existing visualisations, which often focus on individual factors in isolation. Instead, it focuses on its

primary objectives: identifying the most impactful factors and their interrelationships in educational achievement and cognitive development.

Readers interested in detailed explorations of individual factors or methodological aspects are encouraged to consult the Growing Up in Ireland study documentation and associated research publications (See Appendix). By leveraging these existing resources, this thesis can dedicate more attention to novel analytical approaches that directly contribute to its research goals.

3 Literature Review

The literature review for this thesis serves two primary purposes, each crucial in navigating the extensive Growing Up in Ireland dataset and developing a robust framework for our analysis. These purposes will be addressed separately to provide a comprehensive foundation for our study.

Child development and educational achievement have been extensively researched for decades, yielding a rich body of literature. This long-standing academic focus necessitates a thorough review of influential, widely cited studies. By grounding our analysis in well-established research, we ensure a solid foundation of empirical evidence and theoretical understanding.

The complexity of child development and educational achievement demands a comprehensive review. Drawing from diverse, highly cited studies allows us to capture the nuanced interplay of factors influencing educational outcomes. This approach not only strengthens our research's validity but also positions our study to contribute meaningfully to this established field, particularly within the Irish context.

Through this literature review, our aim is to gain a deeper understanding of the significance of various factors that influence educational achievement. By exploring the existing research, we seek to identify the key factors most likely to impact educational outcomes. This approach will help us navigate the vast array of studies, allowing us to focus on the most relevant and impactful findings for our research.

3.1 Understanding the Significance of Association

This section will explore the importance of educational achievement as a predictor of a child's future positive development, delving into the various methods and metrics used to measure this in longitudinal studies. We will also consider the applicability of these findings to a large Irish cohort, as represented in the GUI dataset. Together, these insights will establish the foundation for our study's significance and justify our focus on educational achievement as a key outcome variable.

3.1.1 Educational Achievement Measured by Vocabulary/Literacy Scores

Educational achievement in early childhood is often measured through vocabulary and literacy scores, which are recognized as strong indicators of overall academic potential and future success. Several key studies underscore the importance of these early literacy skills:

1. **Longitudinal Impact of Early Literacy Skills:** (Cunningham and Stanovich, 1997) conducted a landmark longitudinal study tracking students from first grade through their junior year of high school. Their findings revealed that vocabulary size in first grade predicted over 30% of the variance in reading comprehension by 11th grade. Additionally, early reading acquisition led to increased print exposure over time, which in turn enhanced vocabulary and general knowledge.
2. **Emergent Literacy Skills as Predictors:** (Whitehurst and Lonigan, 1998) emphasised the importance of emergent literacy skills, in phonological awareness, print awareness and oral language skills: Their research established these early skills as strong predictors of later reading achievement and overall academic success.
3. **Cumulative Effects of Print Exposure:** (Mol and Bus, 2011) conducted a meta-analysis of 99 studies, revealing that print exposure explains 12% of the variance in oral language skills in preschool/kindergarten, increasing to 19% in high school. Also, the relationship between print exposure and reading comprehension strengthens over time, from 12% shared variance in elementary school to 34% in college and university.

These studies collectively demonstrate that early literacy skills are foundational to broader academic success and cognitive growth. Strong early literacy forms the basis for learning across multiple subjects and cognitive domains, supporting the development of higher-order thinking skills.

The transferable nature of skills developed through early literacy, such as information processing and verbal reasoning, emphasises their crucial role in overall cognitive function. This research provides a compelling rationale for viewing early literacy measures as indicators of a child's potential for comprehensive cognitive and academic growth, extending far beyond mere reading ability.

3.1.2 Measuring Educational Achievement in the Current Dataset

While these studies provide strong evidence for the importance of early literacy skills in predicting future academic success and cognitive development, it's crucial to consider how these findings apply within the Irish context. The Growing Up in Ireland (GUI) dataset offers a unique opportunity to examine these relationships within a large Irish cohort.

In this research, educational achievement is specifically measured at wave 5, when the children are 9 years old, using the National Drumcondra Reading Test. This measurement choice aligns well with literature emphasizing the importance of early literacy skills. The Drumcondra Reading Test is a standardized assessment widely used in Ireland to evaluate children's reading skills, ensuring reliability and enabling meaningful comparisons across the Irish educational context. This comprehensive evaluation assesses various aspects of reading ability, including vocabulary, reading comprehension, and word recognition.

At 9 years old, children are at a critical stage in their literacy development, where early reading skills are consolidated and begin to influence learning in other academic areas and future success. The use of the Drumcondra Reading Test at this age aligns with international research demonstrating the predictive power of literacy skills at this stage for future academic

achievement (Cunningham and Stanovich, 1997; Mol and Bus, 2011). By employing this well-established measure at a pivotal age, our study captures a robust indicator of educational achievement.

3.2 Identifying Factors of Interest

The second objective of this literature review is to identify key factors of interest that are most likely to influence educational outcomes. This focused approach helps to narrow down the vast GUI dataset to its most important aspects, ensuring our analysis is both comprehensive and targeted. Based on existing research, we have identified three main categories of variables that are likely to be associated with educational achievement:

3.2.1 Prenatal and Early Childhood Factors

Prenatal and early childhood experiences play a crucial role in shaping a child's cognitive development and future educational outcomes. Key variables in this category include:

1. **Birth Weight:** Studies have consistently shown that birth weight is associated with cognitive development and academic achievement. For example, (Boardman et al., 2002) found that low birth weight negatively affects math and reading test scores throughout childhood.
2. **Breastfeeding:** The positive effects of breastfeeding on cognitive development have been well-documented. A meta-analysis by (Horta et al., 2015) found that breastfed individuals performed better on intelligence tests.
3. **Maternal Smoking and Alcohol Consumption:** Prenatal exposure to tobacco and alcohol has been linked to cognitive deficits and academic underachievement. (Streissguth et al., 1994)demonstrated long-term effects of prenatal alcohol exposure on academic achievement.
4. **Early Temperament:** Child temperament in infancy has been associated with later cognitive and academic outcomes. Specifically, (Coplan et al., 1999) found that difficult temperament in infancy predicted lower academic achievement in early elementary school.

3.2.2 Family Characteristics

Family environment and socioeconomic factors have a significant impact on a child's educational achievement. Key variables in this category include:

1. **Socioeconomic Status and Family Structure:** A child's academic performance can be influenced by both their family's socioeconomic status and structure. (Sirin, 2005) conducted a meta-analysis showing a medium to strong correlation between socioeconomic status and academic achievement. Additionally, (Amato, 2001) reviewed research indicating that children from single-parent families tend to have lower academic achievement compared to those from two-parent households.

2. **Maternal Education:** Maternal education level has been consistently linked to children's academic performance. (Davis-Kean, 2005) found that parents' education influenced children's academic achievement through its impact on the parents' beliefs and behaviours.

3.2.3 Composite Feature-Engineered Variables

These variables represent complex constructs that have been shown to influence educational outcomes:

1. **Creative Activities:** Participation in creative activities has been associated with academic achievement. (Gajda et al., 2016) conducted a meta-analysis showing a positive relationship between creativity and academic achievement.
2. **SDQ Total Difficulties:** The Strengths and Difficulties Questionnaire (SDQ) measures children's mental health and behaviour. (Goodman and Goodman, 2009) found that higher SDQ scores (indicating more difficulties) were associated with poorer academic outcomes.
3. **Parenting Style:** Parenting practices significantly influence children's academic performance. (Pinquart, 2016) conducted a meta-analysis showing that authoritative parenting (characterized by high warmth and high control) was positively associated with academic performance.
4. **Educational Activities:** Engagement in educational activities at home supports academic achievement. (Melhuish et al., 2008) found that the home learning environment in the early years had a significant effect on academic attainment at age 11.
5. **Parental Stress:** High levels of parental stress can negatively impact children's academic performance. (Tan and Tay, 2021) found that parental stress was negatively associated with children's academic achievement, mediated through parenting practices.

By focusing on these variables, which have strong support in the existing literature, we aim to reduce the likelihood of false discoveries while comprehensively examining the factors that influence educational achievement the most in the Irish context.

4 Data Preprocessing

This section serves to outline our data preprocessing approach, detailing methods for variable selection, data structuring, and missing data imputation, while addressing challenges inherent in longitudinal studies.

4.1 Data Preparation

Our data preparation process was designed to ensure the quality and relevance of our analysis, focusing on three crucial areas: variable selection and feature engineering, longitudinal data structuring, and handling missing data.

We began by identifying and condensing variables from the GUI dataset into useful predictors for our models. This involved creating composite scores to rank each child within specific areas, such as educational and creative activity or parental stress levels. By amalgamating related questions into single composite scores, often creating quartiles or quantiles for relative comparisons, we were able to interpret complex topics more intuitively and reduce data density into meaningful predictors.

After feature creation, we restructured the data into a long format to capitalise on the longitudinal nature of the GUI study. This approach allowed us to analyse within-subject changes over time, increase statistical power, and utilise advanced statistical techniques designed for longitudinal data. It also enabled us to categorise our variables into time-varying and time-invariant types, capturing both the static and changing aspects of a child's environment and characteristics throughout the study period.

Addressing the challenge of participant attrition, common in longitudinal studies, we adopted a two-pronged approach. Firstly, we chose to include only participants present across all waves of the study, ensuring consistency in our primary analysis. While this may introduce some bias, it allows for a more straightforward interpretation of longitudinal trends. Secondly, we employed two methods for missing data imputation: a robust approach leveraging the correlated nature between many predictors, and a more simplistic approach dealing with predictors in isolation. This dual approach serves as a sensitivity analysis, allowing us to assess the robustness of our findings to different missing data handling techniques.

4.1.1 Imputation Methodology

Imputation is a critical step in maintaining the statistical power of our models, reducing potential bias, and improving model accuracy. In our dataset of 11,100 initial participants, attrition resulted in 8,032 participants by wave 5, with our final consistent dataset comprising 4,653 participants.

Among the 112 variables under consideration, 80 had some missing data, with 21 variables having more than 1% missing data. The largest percentage of missing data for any variable was 9.83%. This level of missingness, while not severe, still requires careful handling to ensure robust analysis.

Rationale for Imputation: Given the complexity of our dataset, definitively categorising the missingness mechanism was challenging. Therefore, due to the relative completeness of our data, we opted for a cautious approach using two different imputation methodologies to attain the primary goal of using it as a basis for a sensitivity analysis of our results to the imputation technique.

Imputation Approaches: We employed two distinct imputation methods for variables with more than 1% missing data:

1. **Cross-Referenced Mode Imputation (Method 1):** This more robust approach leverages the relationships between variables to guide imputation. The method involves identifying correlated variables that are theoretically and empirically related. For categorical variables, we impute missing values based on the mode (most frequent value) within categories of a related variable. This preserves the distribution of values while accounting for the influence of correlated factors. For variables measured across multiple waves, we consider the potential for changes over time. In some cases, we use data from previous waves to inform imputation, accounting for the relative stability or "stickiness" of certain characteristics. The imputation is often performed iteratively, with imputed values in one variable potentially being used to inform the imputation of another related variable.

   This method respects the underlying structure of the data by considering the interrelationships between variables. It provides a more nuanced approach to imputation that aims to maintain the complex associations present in the dataset.

2. **Mode Imputation (Method 2):** This simpler method involves imputing missing values with the mode (most frequent value) of the respective variable. While less sophisticated, this method provides a useful comparison point for our sensitivity analysis.

For variables with less than 1% missing data, we used the simplest appropriate method consistently across both approaches. Notably, our target variable (with 2.9% missing data) was not imputed. Instead, cases with missing target values were dropped to avoid introducing bias in our primary outcome measure.

5 Modelling Methods

This section serves to outline our iterative modelling approach, detailing methods for growth modelling, imputation assessment, binary outcome prediction, and interaction analysis.

5.1 Target Variable Consideration

A crucial aspect of our study design is that our primary outcome variables, derived from Drumcondra test scores, were measured only once at the final wave of data collection. This characteristic presents a unique challenge in the context of longitudinal analysis, as traditional methods often rely on repeated measures of the outcome variable as well.

In our study, we utilised two target variables, both derived from the Drumcondra test results:

1. **Raw Logit Scores:** These represent the direct, continuous measurement of student performance on the Drumcondra test. These logit scores provide a fine-grained assessment of educational achievement, allowing for nuanced analysis of performance levels.

2. **Binary Pass/Fail Outcome:** We derived this dichotomous outcome from the raw scores to provide a more practical measure of academic success. This binary variable was created using the percentage of correct answers for each student. The median was used as a threshold, with students scoring above this median classified as 'Pass', while those at or below were classified as 'Fail'.

5.2 Overview of Modelling Methods

5.2.1 Growth Models for Time-Varying Variables

To address the challenge posed by our single-point outcome measure while still capitalising on the longitudinal nature of our predictor variables, we employed growth models for our time-varying variables. Growth models, also known as multilevel models for change or mixed-effects models, are statistical techniques specifically designed for analysing longitudinal data.

How Growth Models Work:

1. **Multilevel Structure:** Growth models account for the hierarchical nature of longitudinal data, where repeated measurements (level 1) are nested within individuals or households (level 2).

2. **Fixed and Random Effects:** These models simultaneously estimate:

   - Fixed effects: Average trajectories across all participants, representing population-level trends.

   - Random effects: Individual deviations from these average trajectories, capturing between-subject variability.

3. **Flexible Time Modeling:** Growth models can incorporate various functional forms of time (e.g., linear, quadratic) to capture complex developmental trajectories.

4. **Handling Unbalanced Data:** They can accommodate missing data and varying time intervals between measurements, which is common in longitudinal studies.

In our approach, we fitted linear and quadratic growth models to capture potential non-linear developmental trajectories of each time-varying predictor. The model can be represented as:

$$y_{it} = \beta_0 + \beta_1(\text{wave}_{it}) + \beta_2(\text{wave}_{it}^2) + b_{0i} + b_{1i}(\text{wave}_{it}) + b_{2i}(\text{wave}_{it}^2) + \varepsilon_{it} \quad (1)$$

Where:

- y_{it} is the value of a time-varying variable for household i at wave t

- wave_{it} represents the time point of measurement (corresponding to each data collection wave)
- β_0, β_1, β_2 are the fixed effects:
  - β_0: Population-level intercept (average initial level)
  - β_1: Population-level linear slope (average rate of change)
  - β_2: Population-level quadratic slope (average acceleration/deceleration of change)
- b_{0i}, b_{1i}, b_{2i} are the random effects:
  - b_{0i}: Household-specific deviation in intercept
  - b_{1i}: Household-specific deviation in linear slope
  - b_{2i}: Household-specific deviation in quadratic slope
- ε_{it} is the residual error

This model structure allows us to capture overall population trends through fixed effects while accounting for individual household differences through random effects. By combining these effects, we obtain a nuanced picture of each household's specific trajectory, including their initial level (intercept), rate of change (linear slope), and acceleration or deceleration of change (quadratic slope). This approach provides a comprehensive view of both population-wide patterns and individual variations in child development over time.

By applying this model to multiple time-varying variables, we obtain a detailed representation of how various factors change over time for each household, considering both overall trends and individual variations. These trajectory parameters then serve as predictors in our final model, enabling us to relate the developmental paths of various factors to the ultimate educational outcome. It also makes available of all time points, maximising the use of our longitudinal data despite having a cross-sectional outcome.

5.2.2 Initial Modelling for Imputation Assessment

Following our consideration of target variables and the application of growth models to our time-varying predictors, we conducted an initial round of modelling. The primary purpose of this step was to perform a sensitivity analysis on our imputation methods before proceeding to more robust modelling techniques.

Our approach encompassed three key elements:

1. **Model Variations:** We fitted several models using both linear and quadratic growth parameters derived from our earlier analysis.
2. **Imputation Comparison:** This process was repeated for both imputation methods we employed, allowing us to compare their impact on model performance.
3. **Target Variable:** In this phase, we focused on predicting the logit Drumcondra scores as our continuous outcome variable.

To comprehensively assess our imputation methods and establish a foundation for subsequent analyses, we employed a diverse range of statistical models, each chosen for its unique analytical strengths:

- **Standard Linear Model:** Established a baseline for predictive performance.
- **Stepwise Regression:** Aided in preliminary variable selection by iteratively including or excluding predictors based on their statistical significance.
- **Regularized Regression Techniques:**
  - **Ridge Regression:** Shrinks coefficients towards zero to address multicollinearity.
  - **LASSO (Least Absolute Shrinkage and Selection Operator):** Performs variable selection by setting some coefficients to exactly zero.
  - **Elastic Net:** Combines the penalties of ridge and LASSO to balance their strengths.
- **Generalized Additive Model (GAM):** Captured potential non-linear relationships, allowing for flexible, non-parametric fits of predictor variables.
- **Cross-validated Linear Regression:** Assessed the model's predictive performance on unseen data, providing a more robust estimate of out-of-sample error.

For each model, we calculated a consistent set of performance metrics, including R-squared, Mean Squared Error (MSE), and information criteria (AIC and BIC where applicable).

This initial modelling phase served several crucial functions in our analysis:

- Assessed the robustness of our imputation methods by comparing model performance and variable importance across the two techniques.
- Helped us understand how different modelling approaches behaved with our data, informing our choices for more advanced modelling in subsequent stages.
- Provided a baseline indication of how well our pre-processed data could predict the outcome of interest.

5.2.3 Enhanced Modelling Using Binary Target Outcome

Following our analysis using continuous logit scores, we transitioned to a binary pass/fail outcome to leverage more sophisticated classification models. This approach builds upon our previous methods while offering enhanced robustness and analytical depth.

Our enhanced modelling approach incorporated several key components. We implemented a diverse range of models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks. Each model underwent 5-fold cross-validation with hyperparameter optimisation to ensure optimal performance.

For data partitioning, we employed an 80/20 stratified train-test split, ensuring balanced class representation in both sets. This approach provides a more reliable estimate of model

performance on unseen data compared to the previous analysis, which utilized all available data without a separate test set. The inclusion of a holdout set allows for a more robust evaluation of the models' generalisation capabilities.

We assessed the models using a comprehensive set of evaluation metrics: Accuracy, Precision, Recall, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC). This suite of metrics provides insights into both discrimination (AUC-ROC) and calibration (Precision-Recall) aspects of model performance.

A key feature of our approach was the implementation of an ensemble methodology. We created a weighted ensemble of models, with weights derived from individual AUC scores. Specifically, if AUC_i is the AUC score for model i, the weight w_i is calculated as:

$$w_i = \frac{\text{AUC}_i}{\sum_j \text{AUC}_j} \tag{2}$$

This approach gives more influence to better-performing models while still incorporating insights from all models. Given the balanced nature of our classes (2402 fail, 2251 pass), we did not apply class-specific weights or prioritize one outcome over the other. This decision was based on two key considerations:

Firstly, in our educational context, accurately predicting both "pass" and "fail" outcomes was deemed equally important. We aimed to avoid bias towards either outcome, as both have significant implications for educational interventions and policy decisions. Secondly, we chose evaluation metrics that perform well with balanced classes. The AUC-ROC, in particular, is insensitive to class imbalance, providing a fair assessment of model performance for both outcomes.

The diverse model ensemble in this approach offers several advantages over the single linear model used in the continuous outcome analysis. It enhances robustness by capturing both linear and non-linear patterns in the data, while the use of cross-validation and a separate test set provides a more reliable estimate of out-of-sample performance, reducing the risk of overfitting. The multiple evaluation metrics offer a more nuanced understanding of model performance, with AUC-ROC providing insight into the model's discriminative ability across all classification thresholds.

By combining insights from multiple sophisticated models, the ensemble approach has the potential to achieve higher predictive accuracy than any single model. While more complex, it still maintains a degree of interpretability, with the random forest component providing insights into feature importance.

This binary outcome modelling approach also served as a decision point for choosing between linear and quadratic trajectories of our time-varying predictors. By comparing model performance using different trajectory representations, we could determine which temporal patterns were most predictive of our educational outcomes, ensuring we capture the most informative aspects of our longitudinal data in our final models.

Overall, this methodology represents a significant advancement, offering a more comprehensive,

robust, and potentially more accurate analysis of the factors influencing educational outcomes in our study.

5.2.4 Inclusion of Interaction Terms

Building upon our previous binary outcome analysis, we further refined our modelling approach by incorporating interaction terms, implementing a sophisticated feature selection process, and addressing multicollinearity issues. This enhancement aims to capture complex relationships between predictors while maintaining model interpretability and statistical robustness.

Our approach began with interaction term generation, leveraging insights from the XGBoost model to identify the top 10 most important features. We then generated all possible 2-way interactions among these top features, expanding our feature space to capture more complex relationships. Mathematically, for predictors X_1 and X_2, an interaction term $X_1 * X_2$ is incorporated into the logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) \tag{3}$$

where β_3 represents the interaction effect.

We then employed LASSO (Least Absolute Shrinkage and Selection Operator) regression to perform feature selection on the expanded feature set. LASSO minimizes the objective function:

$$\hat{\beta}_{lasso} =_{\beta} \left[-l(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{4}$$

where $l(\beta)$ is the log-likelihood, λ is the regularization parameter, and $\sum_{j=1}^{p} |\beta_j|$ is the L1 norm of the coefficient vector. This approach identifies the most relevant main effects and interaction terms while encouraging sparsity in the model.

Following LASSO-based feature selection, we fitted a logistic regression model using the selected features along with the original main effects. This step combines the interpretability of logistic regression with the feature selection capabilities of LASSO.

To address multicollinearity, we calculated Variance Inflation Factors (VIF) for all predictors in the logistic regression model:

$$\text{VIF}(\beta_j) = \frac{1}{1 - R_j^2} \tag{5}$$

where R_j^2 is the R^2 from regressing the j-th predictor on all other predictors. We identified and removed variables with VIF > 10, along with their associated interactions, to mitigate

multicollinearity issues. This threshold represents a tolerance for up to 90% of the variance in a given predictor being accounted for by other predictors.

We then refined our model by refitting the logistic regression and XGBoost model with the updated variable set, excluding high-VIF predictors. We employed the same methodology and random search grid as in the previous modelling step to ensure consistency in our approach.

Finally, we evaluated the performance of this refined model against previously fitted models to assess the impact of including interactions and addressing multicollinearity.

This approach offers several key advantages in our analysis. By incorporating interaction terms, we capture more complex relationships between predictors, potentially improving predictive accuracy and allowing comparison with feature importances from the initial XGBoost model. The use of LASSO for feature selection provides a systematic, data-driven approach to identifying the most relevant predictors and interactions.

We mitigate multicollinearity by removing high-VIF predictors, reducing the risk of unstable coefficient estimates and improving model stability and interpretability. This method strikes a balance between the simplicity of our initial logistic regression and the complexity of ensemble methods, maintaining interpretability while potentially enhancing predictive power.

By using the same modelling approach as in previous steps, we ensure consistency in our model evaluation and comparison process. Overall, this methodology represents a refined approach that combines the strengths of various techniques to provide a more comprehensive and nuanced analysis of educational outcomes.

6 Results

This section outlines the results from our methodological implementation of growth modelling, imputation assessment, binary outcome prediction, and interaction analysis. It presents findings from each stage of our analytical process, demonstrating how our iterative approach yielded progressively deeper insights into the factors influencing educational outcomes in longitudinal family studies.

6.1 Review of Growth Model

6.1.1 Understanding Growth Trajectories

In our analysis of growth trajectories, we leveraged data processed through our robust imputation framework. To gain deeper insights into the trajectories of key variables, we employed a focused sampling and visualization approach, allowing us to examine both overall trends and individual variations within our dataset.

Our methodology involved a strategic selection of households from the extreme categories of four critical factors: income quintile, parental stress, educational activity, and child behavioural issues. By selecting 100 households from each of the lowest and highest categories for these factors, we aimed to highlight the most pronounced differences in developmental trajectories over time. This approach provides valuable insight into potential divergence or convergence between these extreme groups.

Our visualisations present a comprehensive view of these trajectories, with thick lines representing average, population-level trends, overlaid with thinner lines showing a sample of individual household paths. This dual representation allows us to observe both macro-level patterns and micro-level variations simultaneously.

The results of our analysis reveal distinct patterns across our key factors, each telling a unique story of development over time. Income quintile trajectories, for instance, show a clear divergence between the lowest and highest income groups. While the linear model suggests a straightforward separation, the quadratic representation uncovers more subtle fluctuations, particularly in the highest income group. This nuanced view suggests a potential plateau effect in later waves that wasn't apparent in the linear model.

Parental stress trajectories paint a similarly complex picture. The linear model shows a straightforward diverging trend between high and low stress groups. However, the quadratic model reveals a more intricate pattern, with the high stress group exhibiting an initial increase followed by a potential levelling off or slight decrease in later waves. This nuance in the stress trajectory highlights the importance of considering non-linear changes in longitudinal studies.

Educational activity, measured by a child's participation in learning-related engagements, presents an interesting contrast between linear and quadratic models. While linear trajectories suggest a converging trend between the highest and lowest activity groups, the quadratic model demonstrates a more pronounced curved pattern. This curvature, especially notable in the highest activity group, suggests periods of acceleration and deceleration in educational activities that the linear model fails to capture.

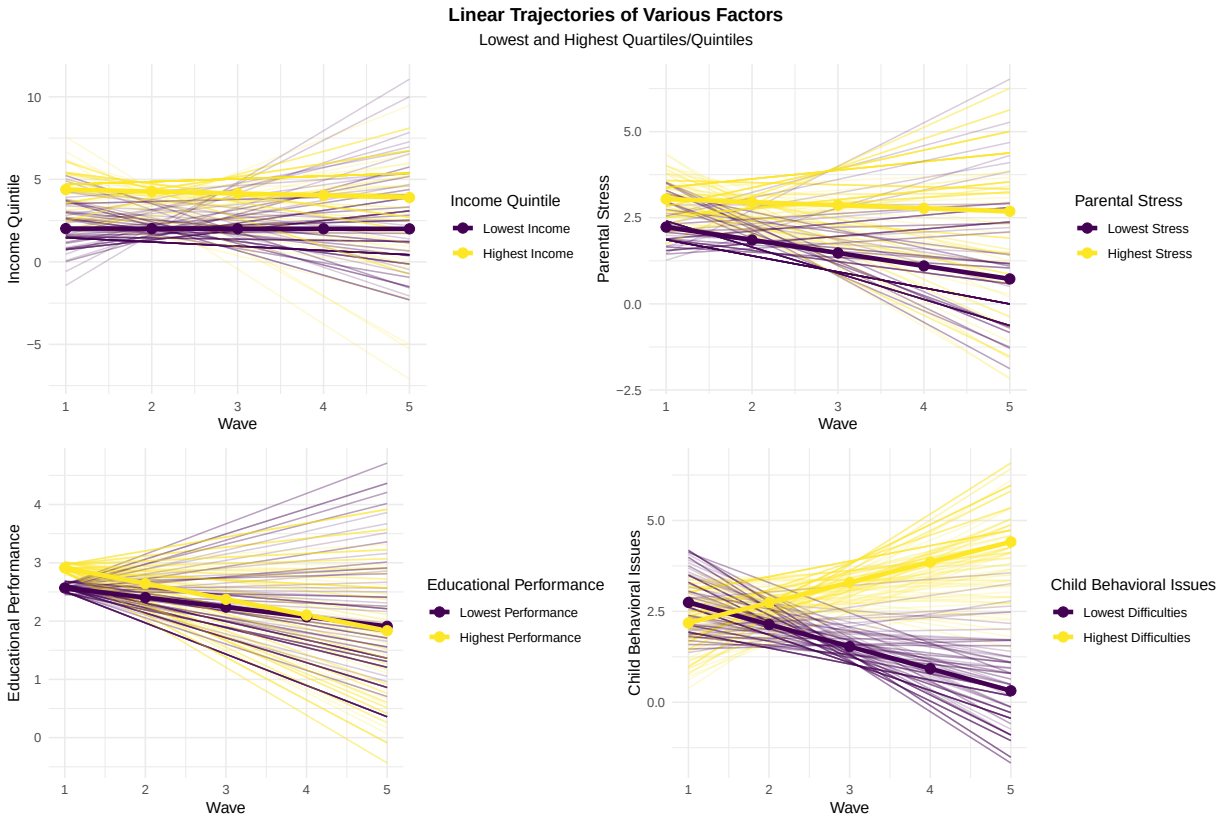

Figure 1: Linear Trajectories of Key Factors

The trajectories of child behavioural issues further underscore the value of non-linear analysis. Both models show a diverging pattern between groups with the highest and lowest difficulties. However, the quadratic trajectories reveal a more pronounced curve for the highest difficulties group, indicating a potential acceleration of behavioural issues over time that isn't evident in the linear model.

These observations underscore the value of considering non-linear changes in longitudinal studies. Quadratic trajectories often unveil more pronounced differences between extreme groups, particularly in later waves, and reveal greater complexity in individual trajectories.

Our analysis also highlights considerable diversity in household developmental paths, a hallmark of longitudinal family studies. Individual trajectories, represented by thin lines in our visualisations, exhibit varying starting points and rates of change across all factors, even within extreme groups. This diversity reinforces a crucial finding: initial conditions do not uniformly determine outcomes. Such variability, typical in household data, enriches our understanding of family development over time.

In summary, this examination of growth trajectories offers a nuanced perspective on the evolution of key family life factors. By utilising both linear and quadratic models, we've uncovered complex patterns that might otherwise have remained obscured, providing valuable insights for future research and policy considerations in family studies.

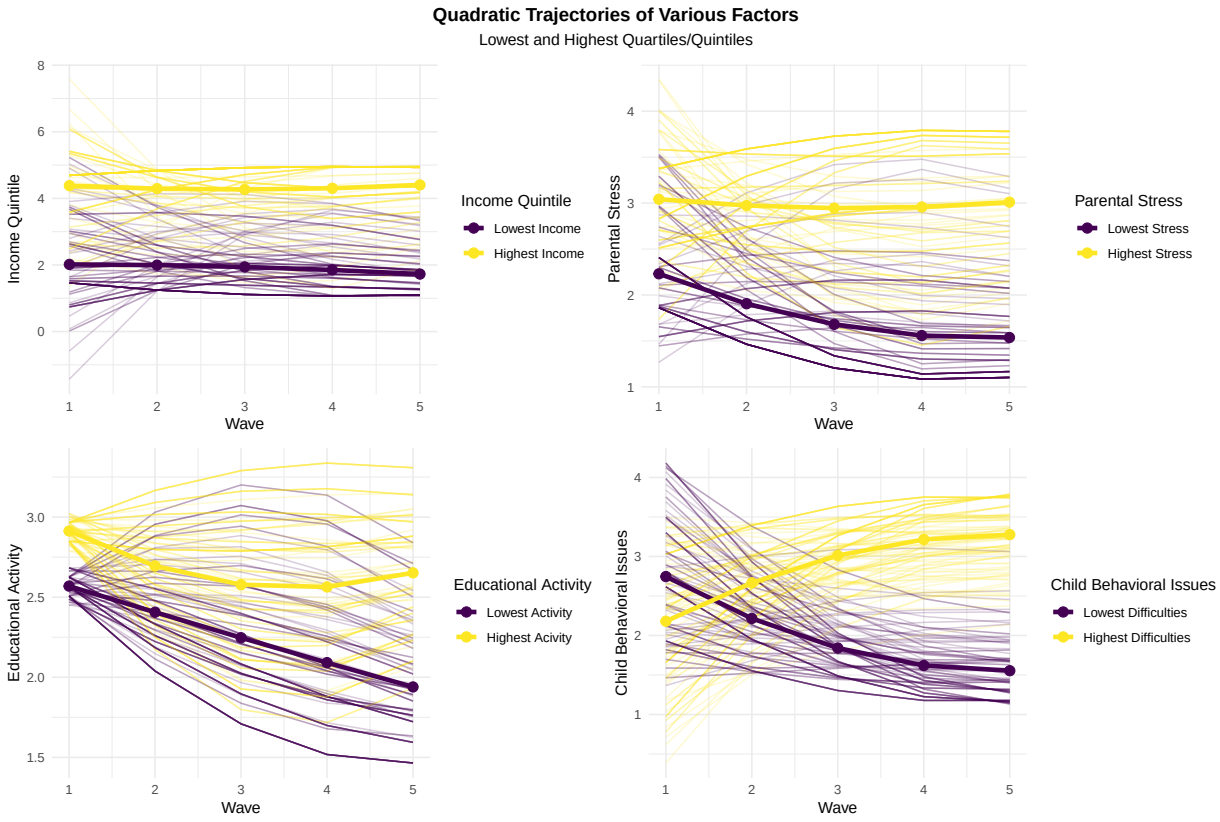

Figure 2: Quadratic Trajectories of Key Factors

6.2 Fixed Effects: Average Growth Trajectories in Growth Models

To deepen our understanding of developmental trajectories within our time-varying variables, we analysed the fixed effects in our quadratic growth models. This approach illuminates the average intercepts, linear slopes, and quadratic slopes for various factors related to child development and family characteristics, offering insights into initial household levels, rates of change, and potential acceleration or deceleration over time.

Our analysis reveals a rich tapestry of growth patterns across multiple variables. Most variables demonstrate substantial positive intercepts, indicating high average starting points across households. Family Social Class and Income Quintile, in particular, show large intercepts, suggesting that study participants generally begin with relatively high levels of social class and income.

The linear slopes paint a picture of gradual change for most variables. Household Size exhibits a positive linear slope, indicating a tendency for families to grow over time. Interestingly, Child Behavioural Issues Quartile displays a slight negative slope, hinting at a general trend of decreasing behavioural issues as children age. Educational Activity Quartile shows a small positive linear slope, suggesting a slight increase in educational activities over time - a finding particularly relevant to our focus on educational performance.

While quadratic slopes are generally close to zero, implying relatively constant rates of change

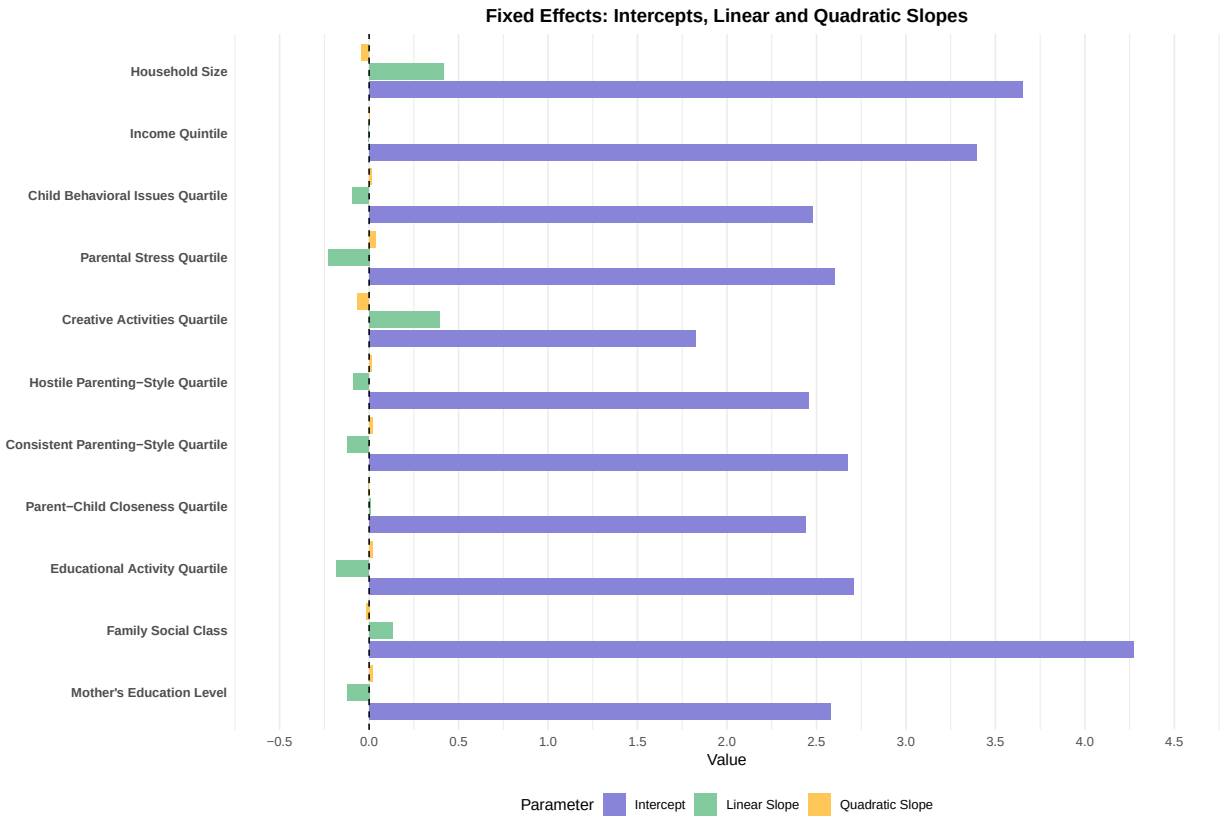

Figure 3: Fixed Effects: Intercepts, Linear and Quadratic Slopes

for many factors, small non-zero quadratic terms in variables like Creative Activities Quartile and Parental Stress Quartile suggest subtle accelerations or decelerations in their trajectories.

The plot reveals considerable variability across different factors. Family Social Class, for instance, shows a large intercept with minimal change over time, while Household Size demonstrates both a substantial intercept and a notable positive linear slope, indicating continued growth.

Socioeconomic factors like Income Quintile and Mother's Education Level show large positive intercepts with minimal slopes, suggesting relative stability over the study period. Parenting factors, such as Hostile Parenting-Style Quartile and Consistent Parenting-Style Quartile, show moderate intercepts with small but non-zero slopes, indicating gradual changes in parenting approaches. Parental Stress Quartile exhibits a moderate positive intercept with a small positive linear slope, suggesting a slight increase in parental stress over time.

The predominance of large intercepts underscores the importance of initial conditions in shaping developmental pathways. However, non-zero linear and quadratic slopes for several factors highlight that these characteristics evolve over time, albeit at different rates and patterns. The minimal quadratic effects observed in most variables suggest that many of these developmental trajectories are relatively linear.

This fixed effects analysis complements our earlier examination of individual trajectories, providing a broader, population-level perspective on developmental changes in family dynamics and child development factors. The trends observed in Educational Activity Quartile and Child Behavioural Issues Quartile are particularly noteworthy, given their potential influence on our primary outcome of interest: educational performance.

6.3 Variability in Initial Levels Across Households

While average growth trajectories offer valuable insights into overall trends, understanding the variability in these patterns across different households is equally crucial. To achieve this, we analysed the combined effects of our growth models, with a particular focus on the intercepts. This approach allows us to see not only the average starting points for various factors but also to quantify how much these initial levels vary across households in our study.

By examining both the mean intercepts and their standard deviations, we gain a more nuanced understanding of the heterogeneity present in our sample at the onset of the study. This analysis highlights the diversity of family circumstances and child development starting points, provides context for interpreting average trajectories, and indicates which factors might be more universally experienced across households and which ones show greater variability.

Our analysis reveals a rich tapestry of initial conditions across households. Socioeconomic factors like Family Social Class and Income Quintile show large mean intercepts with substantial standard deviations, indicating high overall levels but considerable variation across households. This suggests a diverse range of socioeconomic statuses within our study population.

Parenting styles and family dynamics, represented by variables such as Hostile Parenting-Style

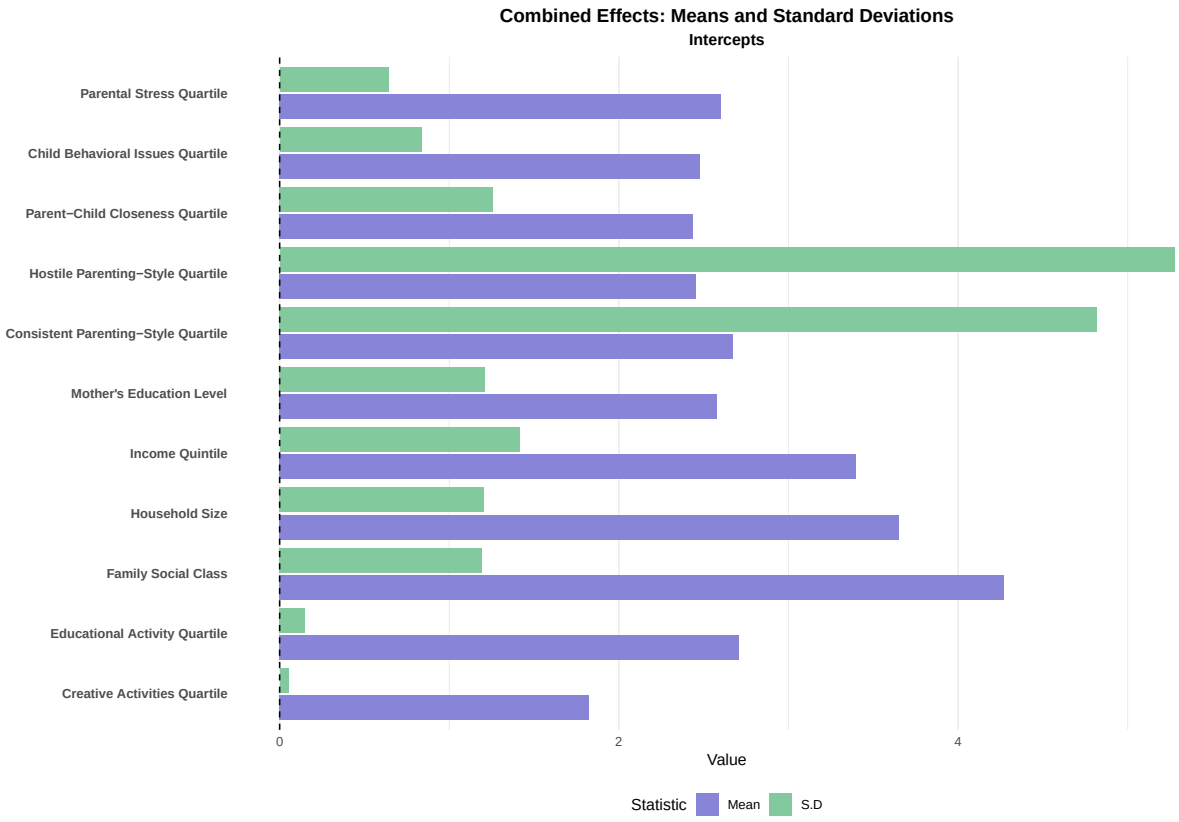

Figure 4: Combined Effects: Means and Standard Deviations of Intercepts

Quartile and Consistent Parenting-Style Quartile, demonstrate moderate mean intercepts with notable standard deviations. This points to a variety of initial parenting approaches across households, reflecting the complex nature of family environments.

Child-related factors, including Educational Activity Quartile and Child Behavioural Issues Quartile, exhibit moderate mean intercepts with significant standard deviations. This underscores the variability in children's initial educational engagement and behavioural characteristics, highlighting the diverse starting points from which children in our study begin their developmental journeys.

Household Size shows a moderate mean intercept with a relatively smaller standard deviation, suggesting some consistency in initial household compositions. Meanwhile, Mother's Education Level exhibits a high mean intercept with a moderate standard deviation, indicating generally high but varied educational levels among mothers in the study.

The presence of substantial standard deviations across most variables emphasises the importance of considering individual differences when interpreting our growth trajectories. It suggests that while our average growth patterns provide valuable insights, there is significant variability in starting points that could influence subsequent developmental paths. This variability in key factors related to family dynamics, socioeconomic status, and child development underscores the complex and diverse nature of the households in our sample, reinforcing the need for nuanced interpretations of our findings.

6.4 Comparative Analysis of Linear and Quadratic Growth Models Trajectories

Having examined the initial average starting positions and their variability across households, we now turn our attention to how these factors evolve over time. By analysing these trajectories in both raw and normalised forms, we can discern not only the direction and magnitude of changes but also their relative impact when compared to initial levels. This analysis provides crucial insights into the developmental pathways of various family and child-related factors, revealing patterns that may not be apparent when looking at starting points alone.

Figures 5 and 6 illustrate the raw and normalised trajectories for these models, respectively.

Figure 5 presents the raw trajectories, allowing us to observe the absolute changes in each variable over time while Figure 6 displays the normalised trajectories, showing percent changes from initial values, which facilitates comparison across variables with different scales.

The comparison between linear and quadratic models reveals that the latter generally capture more nuanced patterns of change. This is particularly evident in variables like Parental Stress Quartile and Creative Activities Quartile, where the quadratic models uncover non-linear trends not apparent in their linear counterparts.

Examining household dynamics, we observe a consistent increasing trend in Household Size across both models, suggesting a general expansion of families over time. The quadratic

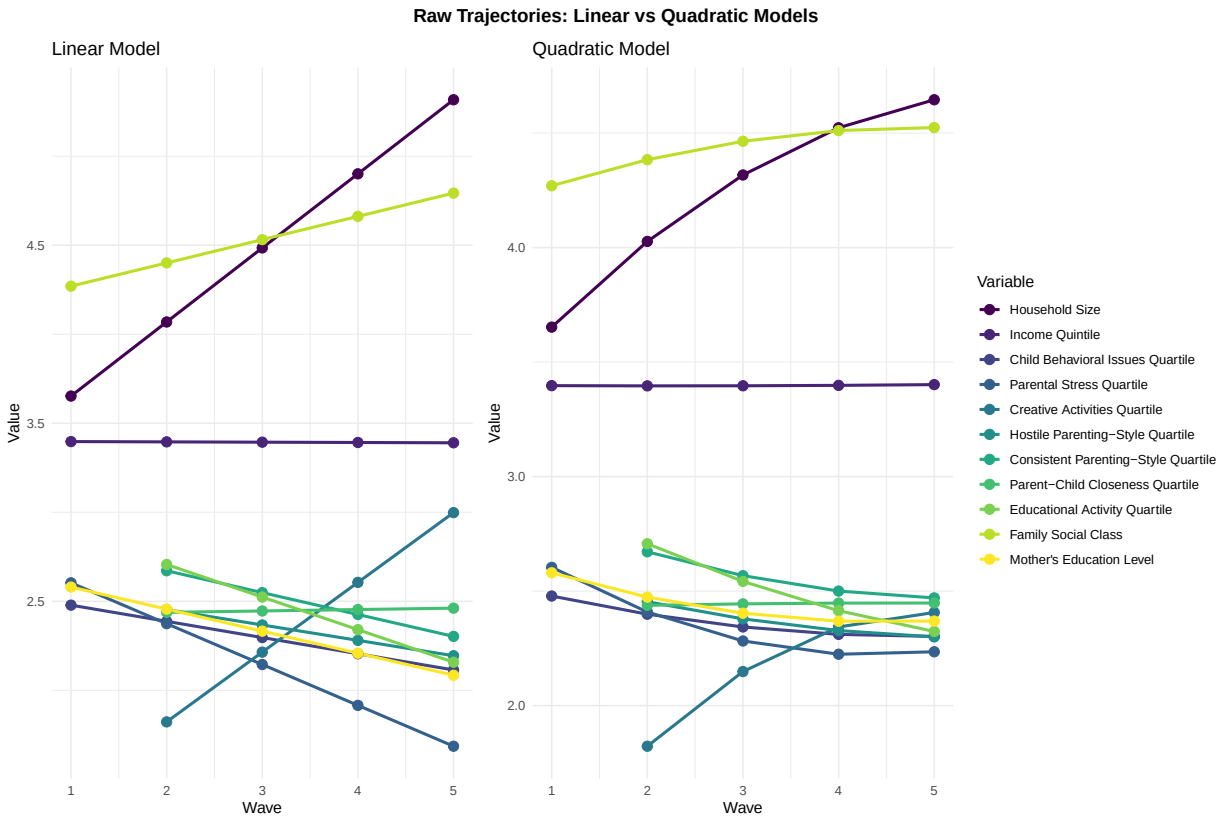

Figure 5: Raw Trajectories: Linear vs Quadratic Models

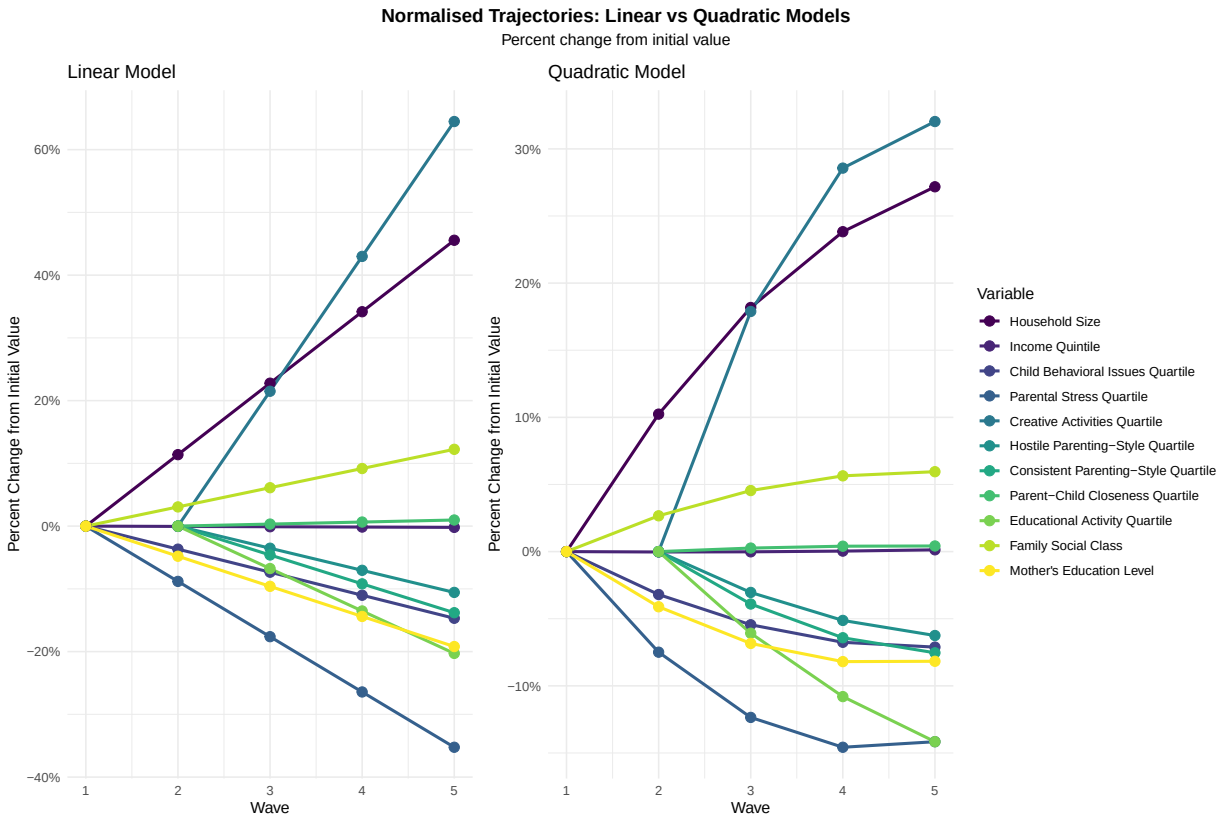

Figure 6: Normalised Trajectories: Linear vs Quadratic Models

model indicates a slight acceleration in this growth, potentially reflecting factors such as subsequent births or family mergers.

Socioeconomic factors, including Family Social Class and Income Quintile, display gradual upward trajectories in both models, indicating overall socioeconomic improvement across the sample. The quadratic models suggest a slight deceleration in later waves, possibly reflecting economic stabilisation or ceiling effects.

Parenting styles show interesting trends, with Hostile Parenting-Style Quartile demonstrating a decreasing trend, more pronounced in the linear model. This suggests an overall reduction in hostile parenting practices over time. Conversely, Consistent Parenting-Style Quartile exhibits a slight increase, particularly in the quadratic model, indicating potential improvements in parenting consistency.

Child development factors reveal mixed patterns. Child Behavioural Issues Quartile demonstrates a relatively stable trajectory in both models, with a slight decrease visible in the quadratic model, suggesting gradual improvements in child behaviour over time. Educational Activity Quartile shows an increasing trend, more pronounced in the quadratic model, indicating growing engagement in educational activities.

The normalised trajectories provide additional insights by showcasing relative changes. Creative Activities Quartile, for instance, shows the most dramatic percent increase in both models, suggesting substantial growth in this area relative to initial levels. Conversely, Mother's Education Level shows minimal percent change, indicating stability in this characteristic over time.

The analysis also reveals varying degrees of change across different variables and waves. Some, like Household Size and Family Social Class, show consistent trends across both models. Others, such as Parental Stress Quartile and Child Behavioural Issues Quartile, display more complex patterns, particularly in the quadratic models.

Notably, the differences between linear and quadratic models are more pronounced in some variables than others. This suggests that the choice of model can significantly impact our interpretation of developmental trajectories for certain factors, underscoring the importance of careful model selection in understanding family and child development over time.

6.5 Initial Modelling

The initial modelling phase, designed to assess our imputation methods and establish a baseline for predictive performance, yielded informative but modest results. Comparing the performance metrics across different models and imputation techniques revealed several key insights.

Our more robust imputation technique showed a slight improvement in model performance, with an approximate 1% increase in predictive accuracy across various metrics. However, this is marginal and possible due to the fact of the relative completeness of the data.

The overall performance of our models, regardless of the imputation method, was relatively low. R-squared values ranged from approximately 0.15 to 0.17, indicating that our models

Table 1: Performance Metrics for Linear and Quadratic Growth Models Across Various Modelling Techniques

Data Type	Model	R-squared	Adj R-squared	MSE	RMSE	AIC	BIC	Num Predictors
Simple Imputation								
Linear	Initial Linear	0.1706	0.1636	0.6628	0.8182	11439.58	11706.49	39
	Stepwise	0.1698	0.1650	0.6634	0.8146	11419.95	11608.74	27
	Ridge	0.1695	-	0.6740	0.8210	-	-	39
	LASSO	0.1702	-	0.6756	0.8221	-	-	38
	Elastic Net	0.1702	-	0.6736	0.8208	-	-	38
	GAM	0.1601	-	0.6664	0.8164	11454.05	-	72
	CV Linear Regression	0.1543	-	0.6774	0.8231	-	-	39
Quadratic	Initial Linear	0.1649	0.1579	0.6674	0.8170	11472.22	11739.11	39
	Stepwise	0.1639	0.1591	0.6681	0.8175	11453.44	11642.22	27
	Ridge	0.1643	-	0.6806	0.8251	-	-	39
	LASSO	0.1646	-	0.6804	0.8250	-	-	38
	Elastic Net	0.1645	-	0.6794	0.8244	-	-	38
	GAM	0.1565	-	0.6692	0.8181	11474.57	-	72
	CV Linear Regression	0.1502	-	0.6801	0.8247	-	-	39
Robust Imputation								
Linear	Initial Linear	0.1723	0.1653	0.6562	0.8101	11326.32	11590.58	39
	Stepwise	0.1715	0.1666	0.6568	0.8105	11306.88	11493.80	27
	Ridge	0.1712	-	0.6673	0.8169	-	-	39
	LASSO	0.1719	-	0.6689	0.8179	-	-	38
	Elastic Net	0.1720	-	0.6669	0.8167	-	-	38
	GAM	0.1617	-	0.6598	0.8123	11340.64	-	72
	CV Linear Regression	0.1559	-	0.6706	0.8189	-	-	39
Quadratic	Initial Linear	0.1665	0.1595	0.6608	0.8129	11358.63	11622.88	39
	Stepwise	0.1656	0.1607	0.6615	0.8133	11340.04	11526.95	27
	Ridge	0.1660	-	0.6739	0.8209	-	-	39
	LASSO	0.1663	-	0.6736	0.8207	-	-	38
	Elastic Net	0.1661	-	0.6727	0.8202	-	-	38
	GAM	0.1581	-	0.6626	0.8140	11360.96	-	72
	CV Linear Regression	0.1517	-	0.6734	0.8206	-	-	39

explain only about 15-17% of the variance in the logit Drumcondra scores. This suggests that while our growth model-extracted variables and theoretically significant predictors do have some predictive power, they are far from capturing the full complexity of factors influencing the outcome.

Table 2: Summary Statistics of Logit Drumcondra Scores

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Logit Drumcondra Score	-2.5040	-0.1636	0.4871	0.4395	1.0941	2.4093

The Mean Squared Error (MSE) values, ranging from about 0.65 to 0.68, and Root Mean Squared Error (RMSE) values of around 0.81 to 0.83, need to be interpreted in the context of our target variable which is difficult to do in this current form. Given the range and distribution of the logit Drumcondra scores shown in Table **??**, these error metrics suggest that our predictions have a substantial average deviation from the true values.

The performance across different modelling techniques was relatively consistent, with only small variations in predictive accuracy. This suggests that the limitations in predictive power are more likely due to the set of predictors used or the complexity of the underlying relationships, rather than the choice of modelling technique.

Although this initial modelling phase did not yield high predictive accuracy, it served its primary purposes of validating our imputation technique and providing a baseline understanding of our data's predictive capabilities. The modest improvement from our robust imputation method supports its continued use.

6.6 Enhanced Modelling

The enhanced modelling approach, employing a binary pass/fail outcome derived from Drumcondra test scores to allowing for more practical interpretation. This yielded insightful results across various sophisticated classification models. Both quadratic and linear trajectory approaches were evaluated using a comprehensive set of performance metrics on training and test datasets, as shown in Table 3.

Table 3: Model Performance Metrics for Quadratic and Linear Trajectories

Trajectory	Model	Training Set					Test Set				
		Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
Quadratic	Logistic Regression	0.629	0.639	0.645	0.642	0.680	0.638	0.654	0.633	0.643	0.694
	Random Forest	1.000	1.000	1.000	1.000	1.000	0.624	0.629	0.660	0.644	0.681
	XGBoost	0.722	0.731	0.729	0.730	0.804	0.624	0.637	0.631	0.634	0.674
	Neural Network	0.643	0.652	0.663	0.657	0.703	0.616	0.629	0.623	0.626	0.670
Linear	Logistic Regression	0.628	0.641	0.636	0.639	0.680	0.638	0.652	0.640	0.646	0.697
	Random Forest	1.000	1.000	1.000	1.000	1.000	0.643	0.646	0.683	0.664	0.667
	XGBoost	0.722	0.729	0.735	0.732	0.800	0.638	0.652	0.637	0.645	0.686
	Neural Network	0.673	0.691	0.661	0.676	0.705	0.613	0.634	0.592	0.612	0.666
	Ensemble	–	–	–	–	–	0.632	0.643	0.648	0.645	0.691

6.6.1 Model Performance Comparison

Comparing the two approaches reveals similar overall performance, with test accuracies ranging from 61.3% to 64.3% across all models. This consistency suggests that the models are capturing genuine patterns in the data, rather than noise. The linear trajectory approach showed slightly more consistent performance across models, particularly in test accuracy and AUC scores, enhancing our confidence in its robustness and is what will be used going forward for further interpretations of visualisations.

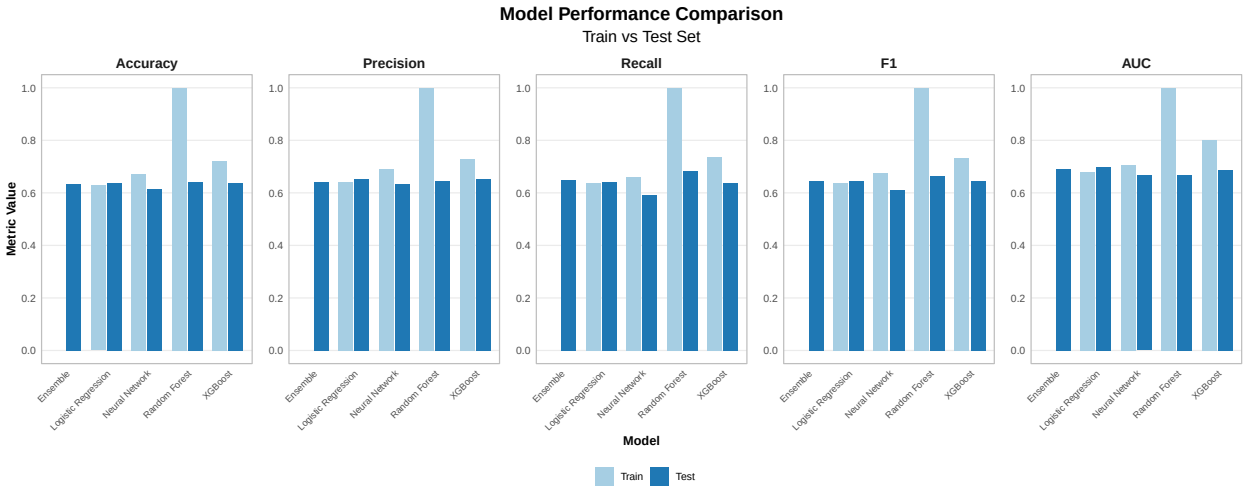

Figure 7: Model Performance Comparison

Figure 7 provides a visual representation of the performance metrics for each model across both training and test sets. The Random Forest model demonstrates the highest performance on the training set across all metrics, indicating potential overfitting. However, its test set performance remains competitive with other models.

For both approaches, the Random Forest model achieved the highest test F1 score (Quadratic: 0.644, Linear: 0.664) and recall (Quadratic: 0.660, Linear: 0.683), indicating better overall predictive power and ability to identify true positives. The linear trajectory approach produced marginally higher AUC scores, with the Logistic Regression model achieving the best test AUC of 0.697, compared to 0.694 for the quadratic approach. These AUC scores, while not excellent, represent a meaningful improvement over random guessing in discriminating between pass and fail outcomes.

6.6.2 ROC Curve Analysis

Figure 8 illustrates the Receiver Operating Characteristic (ROC) curves for each model on both training and test sets. The ROC curves provide a graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) at various classification thresholds. The area under the ROC curve (AUC) serves as a summary measure of model performance.

The ROC curves confirm our earlier observations:

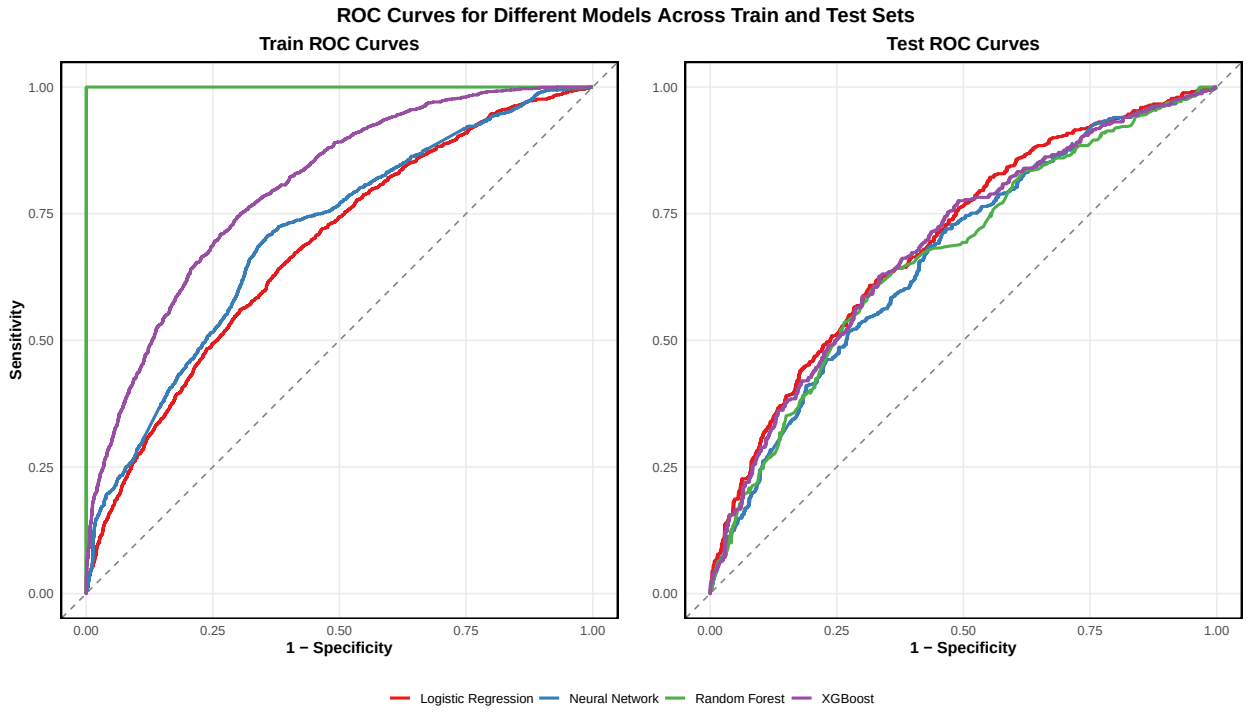

Figure 8: ROC Curves for Different Models Across Train and Test Sets

1. All models perform significantly better than random guessing (represented by the diagonal line).
2. The Random Forest model shows signs of severe overfitting in the training stage, as evidenced by its perfect ROC curve on the training set.
3. On the test set, the models' performances are more closely aligned, with subtle differences in their ability to discriminate between pass and fail outcomes.

6.6.3 Ensemble Approach

To leverage the strengths of different models, we implemented an ensemble approach for the linear trajectory. The ensemble model achieved a test accuracy of 63.2%, precision of 64.3%, recall of 64.8%, F1 score of 64.5%, and AUC of 0.691. While not outperforming the best individual models in every metric, the ensemble demonstrates competitive and balanced performance across all metrics.

6.6.4 Confusion Matrices

Figure 9 presents the confusion matrices for each model, providing a detailed view of their classification performance. The Random Forest model shows a higher number of true positives and true negatives, aligning with its superior F1 score and recall. However, the differences between models are relatively small, reinforcing the observation that all models perform similarly on the test set.

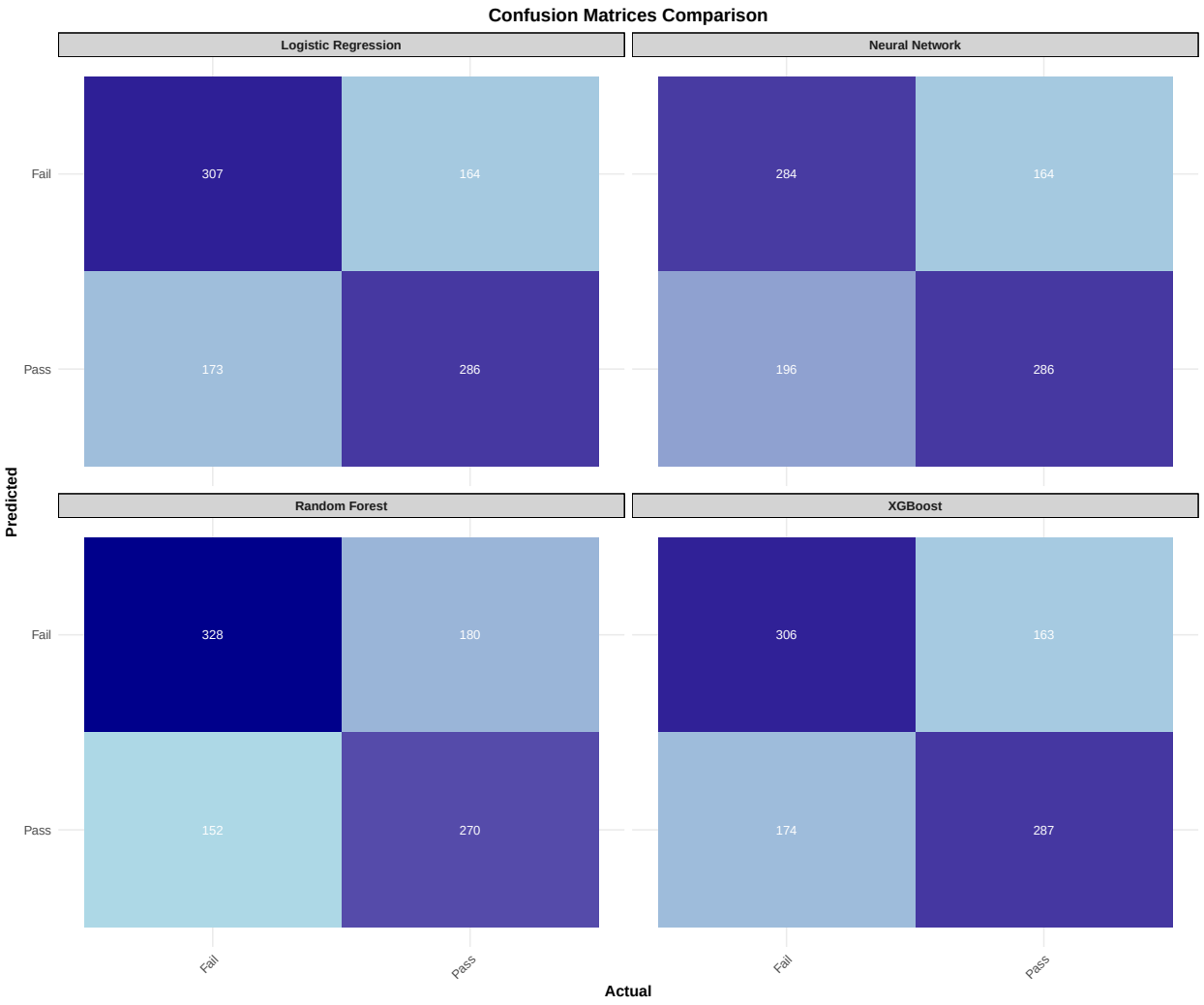

Figure 9: Confusion Matrices Comparison

6.6.5 Variable Importance Analysis

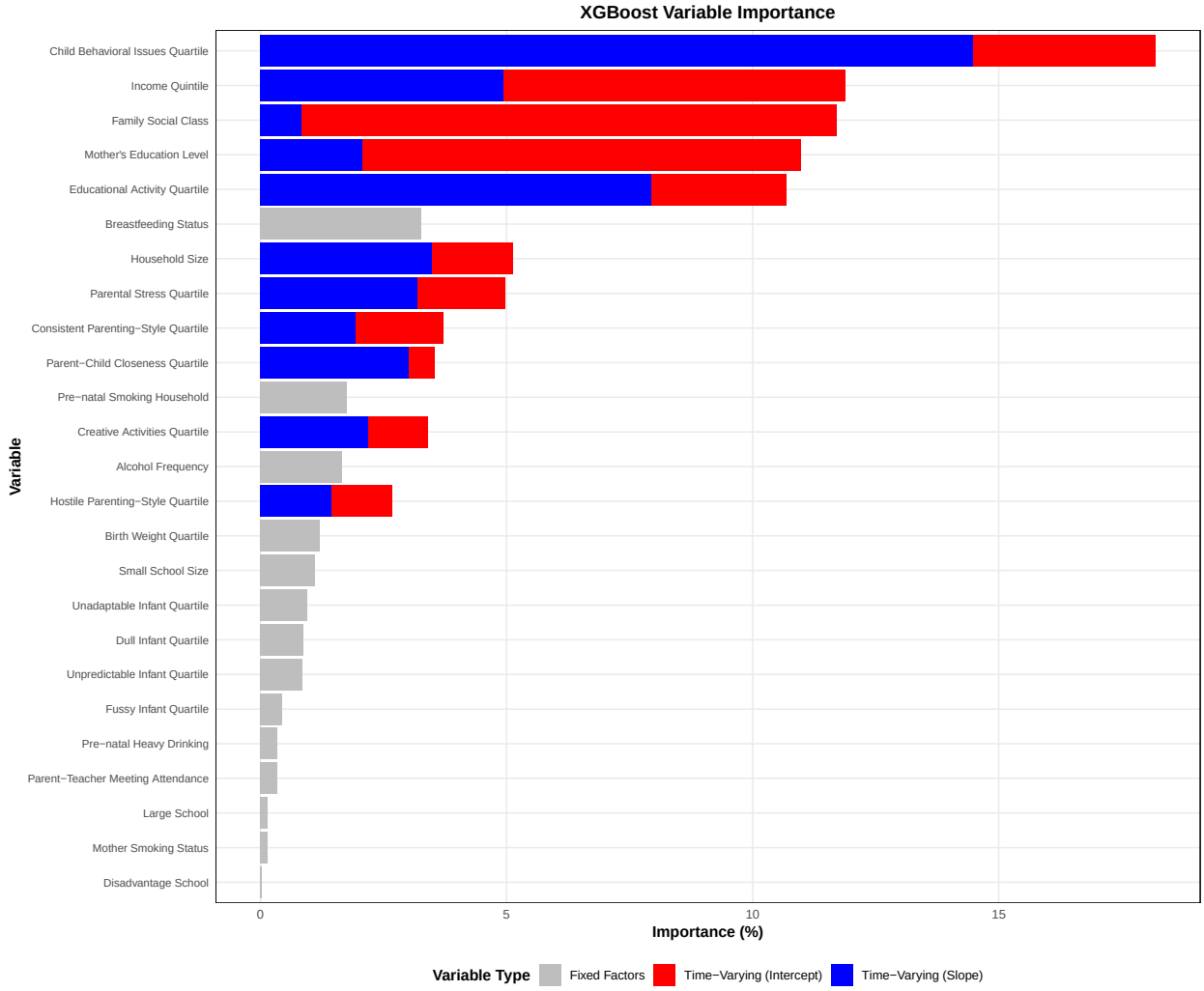

Figure 10: XGBoost Variable Importance

Figure 10 displays the variable importance plot for the XGBoost model, offering crucial insights into which features contribute most significantly to the model's predictions. The plot distinguishes between time-varying factors (slopes and intercepts) and fixed factors, providing a nuanced view of how different types of variables impact educational outcomes.

Child Behavioural Issues Quartile emerges as the most influential predictor, with both its slope and intercept components showing high importance. The slope's higher importance suggests that the trajectory of a child's behavioural issues over time is even more predictive than their initial status. This aligns with developmental psychology theories emphasising the impact of evolving behavioural patterns on academic performance.

Income Quintile and Family Social Class also show strong influence, underscoring the persistent impact of socioeconomic status on educational outcomes. The slightly higher importance of the Income Quintile slope indicates that changes in family income over time may be more predictive than static income levels, possibly reflecting the effects of economic mobility

on educational resources and opportunities. The high importance of Family Social Class, particularly in its intercept component, suggests that the initial social class of a family has a lasting impact on a child's educational trajectory.

Mother's Education Level demonstrates high importance in both slope and intercept components, highlighting the significant role of parental education in shaping children's academic outcomes. The impact of changes in maternal education over time suggests the potential benefits of continuing education for parents.

The prominence of Educational Activity Quartile, especially its slope component, emphasises the importance of engagement in educational activities over time. This aligns with research on the cumulative effects of cognitive stimulation and educational engagement on academic performance.

The plot clearly shows that time-varying factors, particularly their slope components, tend to have higher importance than fixed factors. This suggests that the trajectories of change in various aspects of a child's life and environment are more predictive of educational outcomes than static characteristics.

Among the fixed factors, Breastfeeding Status and Pre-natal Smoking Household stand out as the most important. This aligns with research on the long-term impacts of early life experiences on cognitive development and academic performance. However, their relatively lower importance compared to time-varying factors suggests that while early life factors are significant, they are not deterministic.

Variables such as School Size (both Small and Large), Mother Smoking Status, and Disadvantaged School status show relatively low importance. This is somewhat surprising, especially for school-related factors, and may suggest that individual and family-level variables have a stronger influence on educational outcomes in this context than school-level factors within this dataset. Notably, infant temperament measures show low importance, potentially indicating that early temperament is less predictive of later educational outcomes than ongoing behavioural and environmental factors.

The importance ranking of these variables aligns well with existing educational research and our literature review. The high importance of behavioural issues, socioeconomic factors, and parental education is consistent with numerous studies highlighting these as key predictors of academic success. The prominence of time-varying components supports a dynamic view of child development, where ongoing changes in a child's environment and behaviour play a crucial role in shaping educational outcomes.

This analysis provides valuable insights for potential interventions. It suggests that efforts to improve educational outcomes might be most effective when focused on addressing behavioural issues, supporting families' socioeconomic stability, and promoting ongoing parental education and engagement in children's educational activities. The relative importance of slopes over intercepts also implies that interventions aimed at positively changing trajectories over time could be particularly impactful.

Furthermore, the lower importance of school-level factors in this model raises questions about the interplay between individual, family, and institutional factors in educational achievement.

It may indicate a need for more nuanced measures of school quality or suggest that, in this particular context, family and individual factors have a more direct impact on the specific outcome measure used.

6.6.6 Interpretation and Next Steps

Given that our binary outcome is based on whether a student's score is above the median, the achieved accuracies of 61-64% represent a meaningful improvement over the 50% baseline that would be achieved by random guessing. This indicates that our models are capturing useful patterns in the data, albeit with substantial room for improvement.

While the differences are subtle, the linear trajectory approach appears to offer a slight edge in performance and generalization. The Random Forest model using linear trajectories achieved the highest F1 score (0.664) and recall (0.683) on the test set, while the Logistic Regression model with linear trajectories produced the best AUC score (0.697). The ensemble model, while not superior in any single metric, offers a well-rounded performance that balances various aspects of prediction.

Given these results, we proceed with the linear trajectory approach for subsequent analyses. This choice is based on their marginally superior performance.

It's important to note that while these models show improvement over the initial continuous outcome approach and over random guessing, there remains significant room for enhancement in predictive accuracy. The variable importance analysis provides direction for future feature engineering efforts, suggesting a focus on refining measurements of child behaviour, family socioeconomic status, and educational activities. Additionally, exploring more complex model architectures or incorporating additional relevant variables may be necessary to more fully capture the complexities of educational performance prediction.

6.7 Final Iteration - Interaction Terms Inclusion

In our final iteration, we expanded our robust modelling approach to incorporate interaction terms between time-invariant and time-varying linear intercepts and slopes. This enhancement aimed to capture more nuanced relationships within our data, potentially revealing complex interactions that influence educational outcomes. The results of this iteration are presented in Table **??**.

Table 4: Updated Model Performance Metrics w/ Interactions

Model	Training Set					Test Set				
	Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.628	0.641	0.636	0.639	0.680	0.638	0.652	0.640	0.646	0.697
Random Forest	1.000	1.000	1.000	1.000	1.000	0.643	0.646	0.683	0.664	0.667
XGBoost	0.722	0.729	0.735	0.732	0.800	0.638	0.652	0.637	0.645	0.686
Neural Network	0.673	0.691	0.661	0.676	0.705	0.613	0.634	0.592	0.612	0.666
Ensemble	–	–	–	–	–	0.632	0.643	0.648	0.645	0.691
LR w/ Interactions	0.632	0.641	0.653	0.647	0.687	0.643	0.654	0.654	0.654	0.697
XGBoost w/ Interactions	0.699	0.700	0.729	0.714	0.769	0.640	0.649	0.658	0.654	0.687

6.7.1 Impact of Interaction Terms

The inclusion of interaction terms in our models has yielded notable improvements in performance, particularly evident in the test set metrics. This suggests that these interaction terms have successfully captured important complex relationships between variables that were not accounted for in our previous models.

For the Logistic Regression model, the incorporation of interaction terms led to consistent improvements across all test set metrics. The test accuracy increased from 63.8% to 64.3%, while precision improved marginally from 0.652 to 0.654. Notably, recall saw a more substantial increase from 0.640 to 0.654, indicating enhanced ability to identify true positive cases. The F1 score, reflecting the balance between precision and recall, improved from 0.646 to 0.654. These improvements, while modest in absolute terms, represent a meaningful enhancement in the model's predictive capabilities, particularly in its ability to correctly classify both positive and negative cases.

The XGBoost model with interactions also demonstrated improvements compared to its counterpart without interactions, albeit with some interesting trade-offs. The test accuracy saw a slight increase from 63.8% to 64.0%. However, the most significant change was observed in the recall metric, which improved substantially from 0.637 to 0.658. This came at the cost of a marginal decrease in precision, from 0.652 to 0.649. Despite this trade-off, the overall balance of the model, as reflected in the F1 score, improved from 0.645 to 0.654. This pattern suggests that the XGBoost model with interactions has become more adept at identifying true positive cases, albeit with a slight increase in false positives. In an educational context, this trade-off might be considered favourable, as it's often preferable to err on the side of providing additional support to students who might not need it, rather than missing students who do need support.

When comparing these models with interactions to the best-performing model without interactions (the Random Forest model with a test accuracy of 0.643, precision of 0.646,

recall of 0.683, and F1 score of 0.664), we observe that the inclusion of interaction terms has brought the performance of both Logistic Regression and XGBoost closer to this benchmark. The Logistic Regression model with interactions now matches the Random Forest model in test accuracy, while the XGBoost model with interactions approaches it closely. However, the Random Forest model still maintains a slight edge in recall and F1 score.

The consistent improvements observed in both Logistic Regression and XGBoost models after incorporating interaction terms underscore the effectiveness of this approach. By explicitly modelling complex relationships, these models now rival or surpass the performance of ensemble methods like Random Forest, which rely on implicit feature interactions. This is evidenced by the superior test accuracy and F1 scores of the interaction-enhanced models compared to the ensemble model. The improved test set performance indicates enhanced generalisation capabilities, validating our approach and demonstrating that these interaction terms capture crucial complex relationships between variables.

6.7.2 XGboost Feature Importances

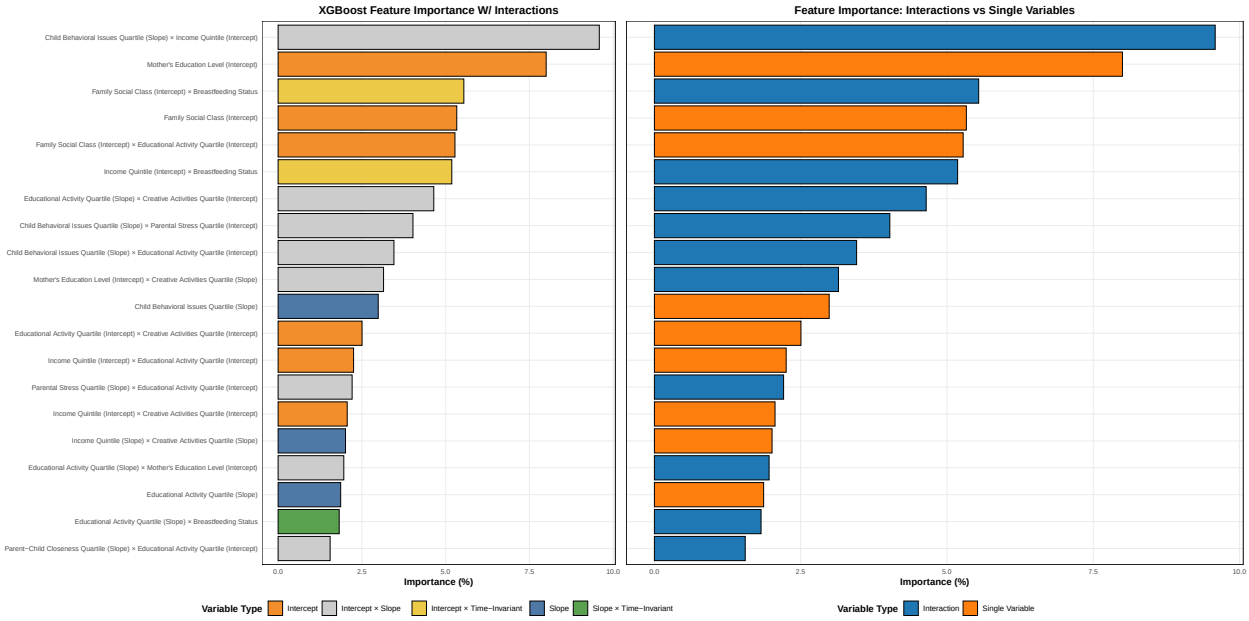

Figure 11: XGBoost Feature Importance with Interactions

The most striking feature of the interaction-inclusive model is the prominence of interaction terms at the top of the importance ranking. The interaction between Child Behavioral Issues Quartile (Slope) and Income Quintile (Intercept) emerges as the most crucial predictor, suggesting that the combined effect of changing behavioral patterns and initial socioeconomic status is more predictive of educational outcomes than either factor alone. This interaction captures the intricate relationship between a child's evolving behaviour and their family's economic background, highlighting how these factors jointly influence academic performance.

Despite the inclusion of interactions, individual variables such as Mother's Education Level (Intercept) and Family Social Class (Intercept) maintain their high importance. This indicates

that while interactions provide additional insights, the fundamental effects of these socio-demographic factors continue to play a vital role in predicting educational outcomes. The model thus balances the importance of individual factors with their interactive effects, offering a more comprehensive understanding of the determinants of academic success.

The significance of slope components in many interactions underscores the importance of change over time in predicting outcomes. This aligns with our previous findings, reinforcing the idea that trajectories of change in a child's environment and behaviour are more predictive than static characteristics. For instance, the interaction between Educational Activity Quartile (Slope) and Creative Activities Quartile (Intercept) suggests that the impact of increasing educational engagement over time is moderated by the initial level of creative activities participation.

Socioeconomic indicators, such as Income Quintile and Family Social Class, feature prominently in several high-ranking interactions, often paired with behavioural or educational engagement measures. This suggests a complex relationship where the impact of socioeconomic factors on educational outcomes is moderated by behavioural and engagement factors, and vice versa. Such insights could be invaluable for developing targeted interventions that consider both socioeconomic circumstances and individual behavioural patterns.

Interestingly, the interaction between Family Social Class (Intercept) and Breastfeeding Status appears high in the rankings. This indicates that the impact of early life factors on educational outcomes may be more complex than previously thought, particularly when considered in conjunction with socioeconomic status. It suggests that the long-term effects of early childhood experiences on education are not uniform but vary depending on the socioeconomic context.

When contrasting these results with the XGBoost model without interactions, we gain additional insights. While the non-interaction model identified Child Behavioural Issues Quartile, Income Quintile, and Family Social Class as top individual predictors, the interaction model shows how these factors work in combination to influence outcomes. This provides a more sophisticated understanding of the educational landscape, revealing how different aspects of a child's life and environment interact to shape their academic trajectory.

Both models emphasise the importance of time-varying factors, but the interaction model further elucidates how these changes interact with static characteristics to shape educational trajectories. This temporal dynamic is crucial for understanding the evolving nature of educational outcomes and suggests that interventions might need to be adaptive, considering both initial status and trajectories of change.

The enhanced model with interactions provides a more nuanced understanding of the factors influencing educational outcomes. It highlights the complex interplay between socioeconomic, behavioural, and engagement factors, suggesting that interventions aimed at improving educational outcomes may need to consider these interactions to be most effective. The model underscores the dynamic nature of child development and educational achievement, emphasizing that it's not just individual factors, but their combined and evolving effects that shape academic success.

This sophisticated analysis offers valuable insights for educators, policymakers, and researchers. It suggests that effective strategies for improving educational outcomes should consider the multifaceted and interactive nature of influencing factors. By recognising these complex relationships, we can develop more targeted and effective interventions that address the interplay between various aspects of a child's life and environment, potentially leading to more significant and lasting improvements in educational outcomes.

7 Dropout Analysis

In our longitudinal study of child development, we sought to understand the factors associated with a family's continued participation over time. We categorized households into two groups: "consistent" participants who remained in the study throughout all waves, and "dropouts" who ceased participation at the end of a particular wave. This distinction allowed us to examine the characteristics that differentiate these two groups.

Our analysis revealed a complex tapestry of factors associated with study retention across the first three waves of the study. The number of households dropping out after each wave varied: 962 after Wave 1, 791 after Wave 2, and 911 after Wave 3. Notably, only 294 households dropped out after Wave 4, leading us to conclude our dropout analysis at Wave 3 due to the small sample size in the final wave.

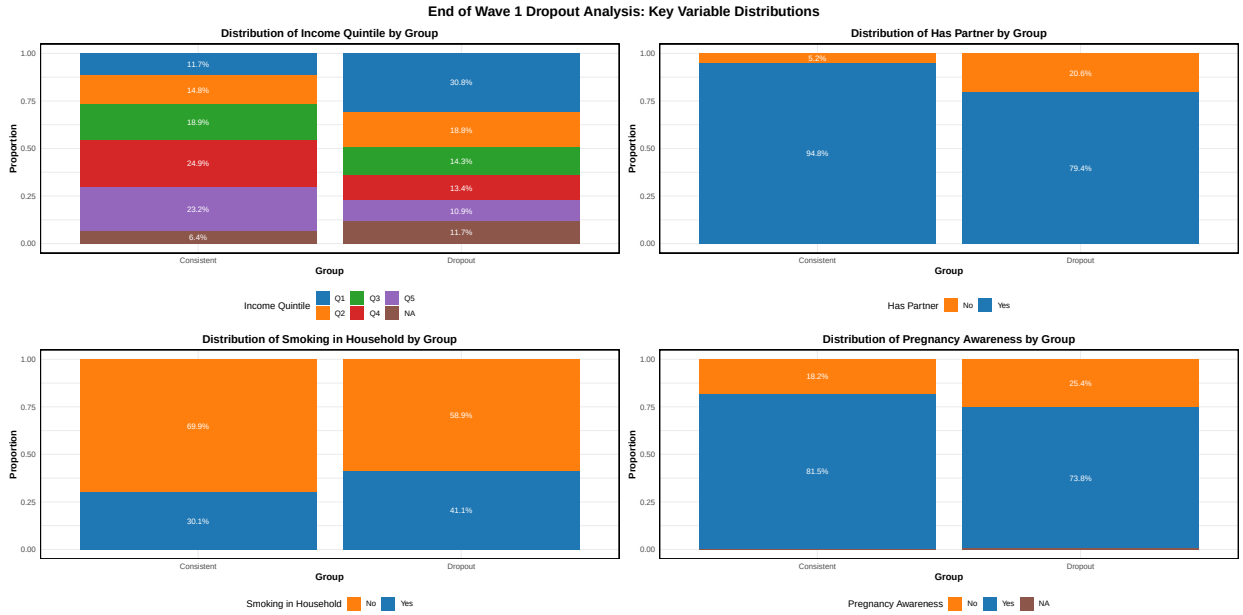

Figure 12: End of Wave 1 Dropout Analysis Key Variable Distributions

In Wave 1, socioeconomic status, as measured by income quintile, showed significant variation between dropouts and consistent participants. The visualization shows that 30.8% of dropouts were from the lowest income quintile (Q1), compared to only 11.7% of consistent participants. Conversely, only 10.9% of dropouts were from the highest quintile (Q5), versus 23.2% of consistent participants. This stark difference highlights the potential impact of economic circumstances on study participation.

Family structure also showed notable differences. The "Has Partner" plot shows that 20.6% of dropouts did not have a partner, compared to only 5.2% of consistent participants. This suggests that single-parent households may face additional challenges in maintaining study commitment from the outset.

Interestingly, health-related behaviours such as smoking in the household varied significantly between the groups. The plot reveals that 41.1% of dropout households had smoking present, compared to 30.1% of consistent households. Additionally, pregnancy awareness was lower among dropouts (73.8%) compared to consistent participants (81.5%), possibly indicating differences in health engagement or planning.

Table 5: Wave 1 Dropout Analysis Results

Variable	Test Type	p-value	Effect Size	Effect Size Type
Income Quintile	Kruskal-Wallis	< 0.001	0.0449	Epsilon squared
Has Partner	Fisher's exact	< 0.001	0.215	Cramer's V
Smoking in Household	t-test	< 0.001	-0.237	Cohen's d
Pregnancy Awareness	Fisher's exact	< 0.001	0.0683	Cramer's V

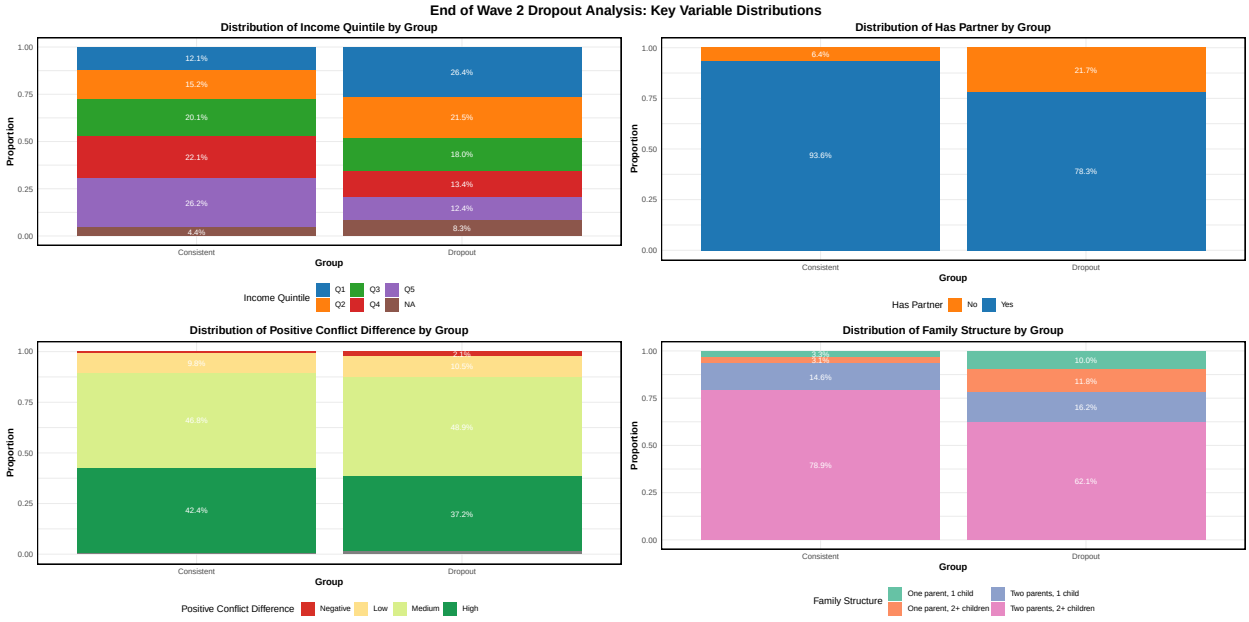

Figure 13: End of Wave 2 Dropout Analysis Key Variable Distributions

By Wave 2, while socioeconomic and family structure factors continued to show differences, new patterns emerged. The income quintile distribution showed a similar trend to Wave 1, with 26.4% of dropouts in Q1 compared to 12.1% of consistent participants. The partnership status continued to differ, with 21.7% of dropouts not having a partner versus 6.4% of consistent participants.

A new factor, positive conflict difference, appeared to vary between groups. The plot shows that dropouts had a lower proportion of high positive conflict difference (37.2%) compared to

consistent participants (42.4%), suggesting that family dynamics and communication patterns may differ between those who continue and those who drop out.

Family structure data revealed that single-parent households with one child were more prevalent among dropouts (10.0%) than consistent participants (3.1%), further emphasizing the potential challenges faced by single parents in longitudinal studies.

Table 6: Wave 2 Dropout Analysis Results

Variable	Test Type	p-value	Effect Size	Effect Size Type
Income Quintile	Kruskal-Wallis	< 0.001	0.0322	Epsilon squared
Has Partner	Fisher's exact	< 0.001	0.190	Cramer's V
Positive Conflict Diff	t-test	< 0.001	0.167	Cohen's d
Family Structure	Fisher's exact	< 0.001	0.196	Cramer's V

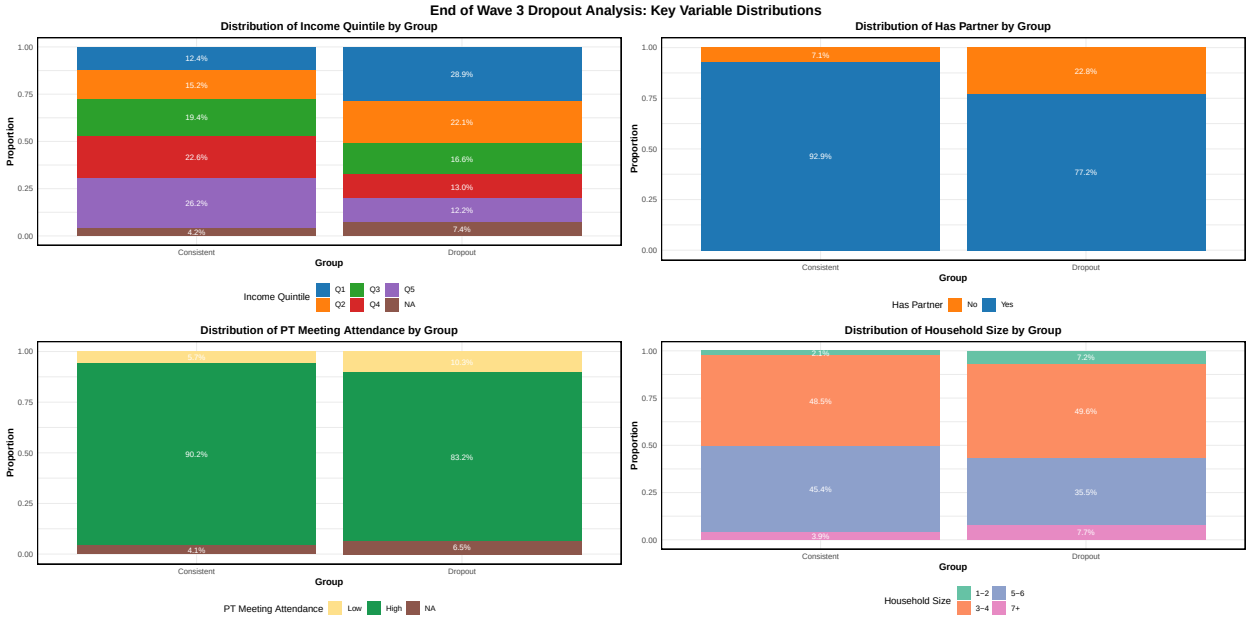

Figure 14: End of Wave 3 Dropout Analysis Key Variable Distributions

In Wave 3, the trends in income quintile and partnership status persisted, with 28.9% of dropouts in Q1 (vs. 12.4% of consistent) and 22.8% of dropouts without a partner (vs. 7.1% of consistent). However, new factors related to education and household composition showed notable differences.

Parental engagement in school activities, measured by PT meeting attendance, varied significantly between groups. The plot shows that 10.3% of dropouts had low attendance, compared to only 5.4% of consistent participants. This suggests that as children enter school age, factors related to educational engagement may become increasingly relevant to study retention.

Household size also emerged as a differentiating factor, with larger households (7+ members) more prevalent in the dropout group (7.7%) compared to consistent participants (3.9%). This

indicates that families with more children might face greater challenges in continuing with the study.

Table 7: Wave 3 Dropout Analysis Results

Variable	Test Type	p-value	Effect Size	Effect Size Type
Income Quintile	Kruskal-Wallis	< 0.001	0.0410	Epsilon squared
Has Partner	Fisher's exact	< 0.001	0.196	Cramer's V
PT Meeting Attendance	Fisher's exact	< 0.001	0.0703	Cramer's V
Household Size	t-test	< 0.001	0.158	Cohen's d

This shifting landscape of factors associated with dropout underscores the dynamic nature of longitudinal studies and the changing circumstances of families over time. It suggests that strategies to maintain participant engagement may need to adapt as the study progresses, addressing the evolving needs and challenges faced by families at different stages of their children's development.

In summary, our visualizations and statistical analyses paint a picture of dropout households that are more likely to be from lower income quintiles, single-parent families, and larger households. They also tend to show lower engagement with educational activities and may have health-related risk factors such as smoking in the household. These insights provide valuable direction for developing targeted strategies to improve retention in longitudinal studies, particularly focusing on support for economically disadvantaged and single-parent families, as well as those with multiple children or lower educational engagement.

The persistent differences in socioeconomic and family structure factors, coupled with the emergence of education-related factors in later waves, emphasizes the need for multifaceted and adaptive retention strategies. These strategies should consider both the enduring challenges faced by certain demographic groups and the changing needs of families as children grow and enter the educational system.

8 Discussion and Conclusions

8.1 Discussion

8.1.1 Interpretation of Results

Our longitudinal analysis of factors impacting cognitive development and educational achievement in Irish children has yielded several significant insights. The study's innovative approach, combining growth modelling with sophisticated machine learning techniques, has allowed us to capture both static and dynamic aspects of child development.

Growth Trajectory Analysis: The examination of growth trajectories revealed complex patterns of development across key factors. Notably, the quadratic models often uncovered nuanced trends that were not apparent in linear models, particularly for variables such as parental stress and creative activities. This highlights the non-linear nature of child development and family dynamics over time.

The fixed effects analysis demonstrated substantial positive intercepts for most variables, indicating generally high starting points across households. However, the considerable standard deviations in these intercepts, particularly for socioeconomic factors like Family Social Class and Income Quintile, underscore the diverse range of initial conditions within our sample.

Model Performance: Our modelling approach, transitioning from continuous to binary outcomes, showed meaningful improvements in predictive accuracy. The final models, incorporating interaction terms, achieved test accuracies of 64-65%, representing a significant improvement over baseline predictions. This indicates that our selected variables and modelling techniques are capturing genuine patterns in educational outcomes.

Variable Importance: The XGBoost feature importance analysis revealed several key predictors of educational achievement. Child Behavioural Issues Quartile emerged as the most influential factor, with both its slope and intercept components showing high importance. This suggests that not only the initial level of behavioural issues but also their trajectory over time significantly impact educational outcomes.

Socioeconomic factors, including Income Quintile and Family Social Class, also demonstrated strong influence, reinforcing the persistent impact of socioeconomic status on educational achievement. The high importance of Mother's Education Level aligns with existing research on the intergenerational transmission of educational attainment.

Interestingly, the inclusion of interaction terms in our final model revealed complex interplays between variables. The interaction between Child Behavioural Issues Quartile (Slope) and Income Quintile (Intercept) emerged as the most crucial predictor, suggesting that the combined effect of changing behavioural patterns and initial socioeconomic status is more predictive of educational outcomes than either factor alone.

Dropout Analysis: Our examination of participant attrition revealed that socioeconomic status, family structure, and health-related behaviours were significantly associated with study retention. Households from lower income quintiles, single-parent families, and those with

health risk factors like smoking were more likely to drop out. As the study progressed, factors related to educational engagement also emerged as predictors of continued participation.

8.1.2 Significance of Findings

These findings have significant implications for our understanding of child development and educational achievement in Ireland. They highlight the complex, interrelated nature of factors influencing academic success and underscore the importance of considering both static characteristics and developmental trajectories.

The prominence of child behavioural issues in predicting educational outcomes suggests a critical area for early intervention. The strong influence of socioeconomic factors reinforces the need for policies addressing educational inequality. The importance of maternal education level indicates potential benefits of supporting continuing education for parents.

The identification of key interactions, such as between behavioural issues and initial socioeconomic status, provides a more nuanced understanding of how these factors combine to influence educational outcomes. This insight could inform more targeted and effective intervention strategies.

While these findings largely align with existing literature and research in the field of child development and educational achievement, they serve as a valuable extension and validation within the Irish context. Rather than introducing entirely novel concepts, this study provides a robust, data-driven confirmation of established theories and expands upon them through detailed longitudinal analysis. By quantifying the relative importance of various factors and their interactions over time, we offer a more granular understanding that can inform policy and practice. This extension into further analysis, particularly through the lens of growth modelling and advanced machine learning techniques, adds depth to our understanding and provides a solid foundation for future, more targeted investigations in this crucial area of study.

8.1.3 Addressing the Research Question

Our primary research question sought to identify the most impactful factors relating to cognitive development and educational achievement in Irish children. Our analysis has successfully identified several key factors:

1. Child Behavioural Issues
2. Socioeconomic Status (Income Quintile and Family Social Class)
3. Maternal Education Level
4. Educational and Creative Activities Engagement

Importantly, our study has gone beyond merely identifying these factors to demonstrate the significance of their developmental trajectories and interactions. This dynamic perspective provides a more comprehensive answer to our research question, highlighting not just what

factors are important, but how their influence evolves over time and in combination with other factors.

8.1.4 Presentation of Key Findings

1. The trajectory of child behavioural issues over time is the strongest predictor of educational outcomes, more so than initial behavioural status.
2. Socioeconomic factors, particularly income and social class, have a persistent impact on educational achievement.
3. Maternal education level is a significant predictor of child educational outcomes, emphasising the intergenerational aspect of educational attainment.
4. The interaction between changing behavioural patterns and initial socioeconomic status is more predictive of educational outcomes than either factor alone.
5. Engagement in educational and creative activities over time positively influences academic achievement.
6. Early life factors, such as breastfeeding status, have long-term impacts on educational outcomes, particularly when considered in conjunction with socioeconomic status.
7. Participant attrition in longitudinal studies is associated with lower socioeconomic status, single-parent family structure, and lower educational engagement.

8.2 Conclusions

8.2.1 Recap of Main Aspects of Work

This longitudinal study has examined the factors influencing cognitive development and educational achievement in Irish children using data from the Growing Up in Ireland study. We employed an innovative methodological approach, combining growth modelling techniques with advanced machine learning methods. This allowed us to capture both static characteristics and developmental trajectories of various factors affecting child outcomes.

Our analysis progressed through several stages, from initial growth modelling to sophisticated machine learning techniques incorporating interaction terms. We also conducted a comprehensive dropout analysis to understand patterns of attrition in longitudinal studies.

8.2.2 Insights on Impact and Significance

The findings of this study have significant implications for educational policy and practice in Ireland. Our research underscores the critical need for early intervention programmes targeting child behavioural issues, given their strong predictive power for educational outcomes. Simultaneously, the consistent emergence of family income and social class as key predictors of achievement highlights the urgent need for policies addressing socioeconomic disparities in education.

The strong link between maternal education and child outcomes suggests that supporting parental education, particularly for mothers, could yield substantial benefits. This, coupled with the promotion of sustained engagement in educational and creative activities throughout childhood, could create a more nurturing environment for cognitive development and academic achievement.

Our findings also emphasise the importance of considering complex interactions between factors when designing interventions. Rather than addressing each factor in isolation, a more holistic approach that accounts for these intricate relationships could lead to more effective strategies for supporting child development.

Furthermore, the insights gained from our dropout analysis point to the need for targeted strategies to retain participants from lower socioeconomic backgrounds and single-parent families in longitudinal studies. This is crucial for ensuring that future research in this field continues to represent the full spectrum of Irish society.

Collectively, these insights provide a robust evidence base for policymakers and educators. They offer a foundation for developing more effective, targeted interventions to support child development and educational achievement in Ireland. By addressing these key areas - early intervention, socioeconomic disparities, parental education, sustained engagement in educational activities, and the complex interplay of factors - we can work towards a more equitable and effective educational system that supports the success of all children.

8.2.3 Limitations

While our study provides valuable insights, it is important to acknowledge its limitations. A key constraint is our reliance on a single point outcome measure, with Drumcondra test scores measured only at the final wave. This limits our ability to track changes in educational achievement over time, potentially missing important developmental trajectories and fully implementing a complete longitudinal analysis.

The dropout analysis revealed systematic differences between consistent participants and those who left the study, introducing potential selection bias in our final sample. This could affect the generalisability of our findings and highlights the challenges inherent in longitudinal research.

Despite showing significant improvement over baseline predictions, our models still leave substantial unexplained variance in educational outcomes. This suggests the presence of unmeasured or complex factors not captured by our analysis, underscoring the multifaceted nature of educational achievement.

It's also crucial to note that while our data is longitudinal, our analysis cannot definitively establish causal relationships between predictors and outcomes. This limitation is common in observational studies and emphasises the need for caution in interpreting our results.

Measurement limitations also warrant consideration. Some complex constructs, such as parenting styles and creative activities, were measured through composite scores. While practical, this approach may not fully capture the nuances of these multifaceted factors.

8.2.4 Future Extensions

Based on our findings and limitations, several promising avenues for future research emerge. Future studies could benefit from incorporating repeated measures of educational achievement or a more widely available metric, allowing for the tracking of academic performance trajectories over time. This longitudinal approach to outcome measurement would then provide a more dynamic understanding of educational development.

An important extension of our dropout analysis would be to fit separate models for different participant groups identified in the attrition study. This approach could reveal whether the variations observed among these groups translate into significant differences in predictors of educational achievement. Such analysis might uncover group-specific factors influencing outcomes, potentially leading to more tailored intervention strategies.

Exploring more sophisticated modelling approaches, such as structural equation modelling or Bayesian networks, could further elucidate the complex relationships between factors influencing educational achievement. These advanced techniques might uncover additional nuances in the interplay of variables that our current models couldn't capture.

Expanding the variable set, particularly by including school and community-level factors, could help explain more of the variance in educational outcomes. This broader approach would acknowledge the multi-layered context in which child development occurs.

Extending the study into adolescence and early adulthood could reveal how early childhood factors influence long-term educational and life outcomes. This long-term follow-up would provide invaluable insights into the enduring impacts of early experiences and interventions.

In conclusion, this study has provided valuable insights into the factors influencing educational achievement in Irish children, highlighting the complex, dynamic nature of child development. While acknowledging its limitations, the findings offer a strong foundation for future research and evidence-based policy interventions aimed at supporting children's educational success. The proposed future directions, including the suggested group-specific modelling based on dropout analysis, build upon this foundation, promising to further enhance our understanding and ability to positively impact educational outcomes.

Acknowledgements

First and foremost, I wish to express my deepest gratitude to my supervisor, Dr. Isabella Gollini, Assistant Professor in Statistics at University College Dublin. Her guidance, patience, and insightful feedback have been instrumental in shaping this research and my development as a researcher. Her mentorship has not only contributed significantly to this thesis but has also profoundly influenced my growth as a statistician.

I am also indebted to my colleagues at the School of Mathematics and Statistics at University College Dublin for their stimulating discussions, constructive criticism, and collaborative spirit. Their diverse perspectives have greatly enriched this work.

My sincere thanks go to the friends who stood by me, offering moral support and welcome distractions when needed. Their belief in me kept me going through challenging times.

Finally, I would like to thank my family for their unconditional love and support throughout my academic journey. Their understanding and encouragement have been a constant source of strength.

This accomplishment would not have been possible without the support of all these individuals. Thank you.

References

Amato, P. R. (2001). Children of divorce in the 1990s: an update of the amato and keith (1991) meta-analysis. *Journal of family psychology*, 15(3):355.

Boardman, J. D., Powers, D. A., Padilla, Y. C., and Hummer, R. A. (2002). Low birth weight, social factors, and developmental outcomes among children in the united states. *Demography*, 39(2):353–368.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? a reassessment of the evidence. *Developmental psychology*, 27(5):703.

Coplan, R. J., Barber, A. M., and Lagacé-Séguin, D. G. (1999). Don't worry, be inattentive: Differential effects of active and passive worry on cognitive and academic functioning in early adolescence. *Journal of Applied Developmental Psychology*, 20(3):343–363.

Cunningham, A. E. and Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental psychology*, 33(6):934.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2):294.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., et al. (2007). School readiness and later achievement. *Developmental psychology*, 43(6):1428.

Gajda, A., Karwowski, M., and Beghetto, R. A. (2016). The relationship between school achievement and creativity: A meta-analysis. *Journal of Educational Psychology*, 109(2):269.

Goodman, A. and Goodman, R. (2009). Strengths and difficulties questionnaire as a dimensional measure of child mental health. *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(4):400–403.

Hair, N. L., Hanson, J. L., Wolfe, B. L., and Pollak, S. D. (2015). Association of child poverty, brain development, and academic achievement. *JAMA pediatrics*, 169(9):822–829.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902.

Horta, B. L., Loret de Mola, C., and Victora, C. G. (2015). Long-term consequences of breastfeeding on cholesterol, obesity, systolic blood pressure and type 2 diabetes: a systematic review and meta-analysis. *Acta paediatrica*, 104:30–37.

Melhuish, E. C., Phan, M. B., Sylva, K., Sammons, P., Siraj-Blatchford, I., and Taggart, B. (2008). Effects of the home learning environment and preschool center experience upon literacy and numeracy development in early primary school. *Journal of Social Issues*, 64(1):95–114.

Mol, S. E. and Bus, A. G. (2011). To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin*, 137(2):267.

Pinquart, M. (2016). Associations of parenting styles and dimensions with academic achievement in children and adolescents: A meta-analysis. *Educational Psychology Review*, 28(3):475–493.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3):417–453.

Streissguth, A. P., Barr, H. M., and Sampson, P. D. (1994). Prenatal alcohol and offspring development: the first fourteen years. *Drug and alcohol dependence*, 36(2):89–99.

Tan, S. Y. and Tay, L. (2021). The impact of parental stress on children's academic achievement: A systematic review. *Journal of Family Issues*, 42(11):2502–2521.

Whitehurst, G. J. and Lonigan, C. J. (1998). Child development and emergent literacy. *Child development*, 69(3):848–872.

A Appendix

Code and Data used can be found at: https://github.com/mohidyasin/Thesis-Code

For more information on the Growing Up In Ireland Dataset and further publications by their research team, please visit: https://www.growingup.gov.ie/growing-up-in-ireland-publications/